UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS / FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇAO EM MODELAGEM COMPUTACIONAL

Thiago Esterci Fernandes

Modelagem da Resposta Imunológica com Aprendizado de Máquina: Validação e Analise de Custo Computacional de PINNs frente a Redes Neurais e Volumes Finitos

Juiz de Fora

2025

Thiago Esterci Fernandes

Modelagem da Resposta Imunológica com Aprendizado de Máquina: Validação e Analise de Custo Computacional de PINNs frente a Redes Neurais e Volumes Finitos

Dissertação apresentada ao Programa de Pós-graduaçao em Modelagem Computacional da Universidade Federal de Juiz de Fora, na área de concentração Modelagem Computacional, como requisito parcial para obtenção do título de Mestre em Modelagem Computacional.

Orientador: Dr. Marcelo Lobosco

Coorientador: Dr. Rodrigo W. dos Santos

Juiz de Fora

2025

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Fernandes, Thiago Esterci.
Modelagem da Resposta Imunológica com Aprendizado de Máquina : Validação e Analise de Custo Computacional de PINNs frente a Redes Neurais e Volumes Finitos / Thiago Esterci Fernandes. -- 2025.
76 f. : il.

Orientador: Marcelo Lobosco
Coorientador: Rodrigo Webber dos Santos
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Modelagem Computacional, 2025.

1. imunologia computaciona. 2. redes neurais informadas por física. 3. modelagem computacional. 4. miocardite infecciosa. 5. edema miocárdico. I. Lobosco, Marcelo, orient. II. Santos, Rodrigo Webber dos, coorient. III. Título.

Thiago Esterci Fernandes

Modelagem da Resposta Imunológica com Aprendizado de Máquina: Validação e Análise de Custo Computacional de PINNs frente a Redes Neurais e Volumes Finitos

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 08 de agosto de 2025.

BANCA EXAMINADORA

Prof. Dr. Marcelo Lobosco - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Leonardo Goliatt da Fonseca
Universidade Federal de Juiz de Fora

Prof. Dr. Vinícius da Fonseca Vieira
Universidade Federal de São João del-Rei

Juiz de Fora, 01/08/2025.

Documento assinado eletronicamente por **Marcelo Lobosco**, **Professor(a)**, em 08/08/2025, às 16:37, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Professor(a)**, em 11/08/2025, às 13:23, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Vinícius da Fonseca Vieira**, **Usuário Externo**, em 13/08/2025, às 10:33, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2526335** e o código CRC **E30467BF**.

Dedico este trabalho a todos que, de alguma forma, fizeram parte da minha trajetória. À minha família, pelo apoio incondicional, em especial aos meus avós Vanderci, Elza e Rui, às minhas tias Lídia e Dorotéia, aos meus pais Angélica e Abel, e à nossa inesquecível Mamamã, por estarem sempre presentes em cada etapa da minha vida. Aos meus amigos de infância Thiago e Vic, companheiros desde as primeiras descobertas, e aos amigos João, Zé, Ítalo e Rafinha, por todos os conselhos e momentos vividos. Aos professores Luiz e Wesley, por terem me abalado profunda e positivamente e também por me ensinarem a importância do pensamento crítico. Ao professor Del, que me ensinou que matemática, como linguagem, tinha seus dialetos e que um professor é um eterno aluno. Aos colegas do programa, em especial João, Guilherme, Eduardo e Alexandro, pelas discussões, aprendizados e pela convivência ao longo desta jornada. Aos professores Bernardo, Flávia e Barbará, que me apoiaram da melhor forma possível e muito além de suas obrigações como coordenadores. Aos meus orientadores Marcelo e Rodrigo, que me acolheram e me guiaram, com muita dedicação e paciência, por uma área nova de conhecimento. A Maíra e a Renata por serem parte imprescindível do funcionamento deste programa. E, finalmente, à minha namorada Ana, por ser ao mesmo tempo minha família, minha colega de trabalho, minha amiga e por ter me acompanhado por todo esse processo.

A todos vocês, dedico este trabalho com gratidão e carinho.

AGRADECIMENTOS

Os autores gostariam de expressar seus agradecimentos à CAPES, FAPEMIG (APQ-02513-22, APQ-02445-24), ao CNPq (308745/2021-3), à FINEP (SOS Equipamentos 2021 AV02 0062/22), ao LNCC (Supercomputador Santos Dumont) e à UFJF pelo financiamento deste trabalho.

São as perguntas que não sabemos responder que mais nos ensinam. Elas nos ensinam a pensar. Se você dá uma resposta a um homem, tudo o que ele ganha é um fato qualquer. Mas, se você lhe der uma pergunta, ele procurará suas próprias respostas. (Kvothe, O Temor do Sábio).

RESUMO

A miocardite infecciosa é uma inflamação do músculo cardíaco frequentemente associada à formação de edema miocárdico, podendo resultar em arritmias, insuficiência cardíaca e morte súbita. Dados epidemiológicos apontam um aumento de 62,2% na incidência global da doença três décadas passadas, totalizando 324.490 óbitos em 2019. Apesar do avanço no entendimento dos mecanismos fisiopatológicos envolvidos, persistem desafios relacionados à modelagem eficiente da resposta imunológica frente à infecção.

Este trabalho tem como objetivo central investigar estratégias para acelerar a simulação da resposta imunológica frente à infecção por meio do uso de redes neurais profundas. Para isso, propõe-se uma análise comparativa entre Redes Neurais Informadas por Física (PINNs) e Redes Neurais convencionais (NNs) na tarefa de modelar a dinâmica espaço-temporal das concentrações de patógenos e leucócitos no tecido miocárdico. Utiliza-se um modelo baseado em equações diferenciais parciais (EDPs), validado previamente na literatura, que descreve os processos de difusão, quimiotaxia, replicação patogênica e recrutamento leucocitário em um meio poroso saturado.

Verificou-se que as PINNs demonstram tendência à suavização excessiva e perda de precisão local em regiões com gradientes acentuados. Já as NNs mostraram excelente capacidade de representar fenômenos localizados e estruturas estacionárias, desde que houvesse densidade amostral adequada. Em termos de desempenho, ambas as abordagens superaram significativamente o Método dos Volumes Finitos (MVF) mesmo em sua versão paralelizada em GPU via CUDA, que apresentou uma aceleração média de 483 vezes. As redes neurais alcançaram aceleração computacional média de 6,66 vezes em relação ao MVF em sua versão GPU, resultado atribuído à possibilidade de avaliação totalmente paralela do domínio espaço-temporal, sem dependência sequencial entre os passos de tempo.

Em um contexto caracterizado por soluções com variações abruptas, cobertura espacial uniforme e dados sintéticos, conclui-se que as redes neurais clássicas apresentaram desempenho superior, conciliando fidelidade à solução de referência e eficiência computacional. Por outro lado, as PINNs mostraram-se mais robustas à escassez de dados, o que reforça seu potencial para aplicações em contextos com menor disponibilidade de dados.

Palavras-chave: imunologia computacional; redes neurais informadas por física; modelagem computacional; miocardite infecciosa; edema miocárdico.

ABSTRACT

Infectious myocarditis is an inflammation of the cardiac muscle often associated with the development of myocardial edema, potentially leading to arrhythmias, heart failure, and sudden death. Epidemiological data indicate a 62.2% increase in the global incidence of the disease over the past three decades, culminating in 324,490 deaths in 2019. Despite advances in understanding the underlying pathophysiological mechanisms, challenges persist in efficiently modeling the immune response to infection.

This study focuses on evaluating deep neural networks as accelerated surrogates for the simulation of the immune response. To this end, a comparative analysis is proposed between Physics-Informed Neural Networks (PINNs) and conventional Neural Networks (NNs) for modeling the spatio-temporal dynamics of pathogen and leukocyte concentrations in myocardial tissue. A model based on partial differential equations (PDEs), previously validated in the literature, is employed to describe the processes of diffusion, chemotaxis, pathogen replication, and leukocyte recruitment in a saturated porous medium.

PINNs tended to smooth the solution excessively, resulting in reduced local accuracy in regions with sharp gradients. In contrast, NNs demonstrated excellent capability in capturing localized phenomena and stationary structures, provided that an adequate sampling density was available. In terms of performance, both neural network approaches significantly outperformed the Finite Volume Method (FVM), even in its GPU-parallelized version using CUDA, which achieved an average acceleration of 483 times. The neural networks reached an average computational acceleration of 6,66 times when compared to the GPU Version of FVM, a result attributed to their ability to evaluate the entire spatiotemporal domain in parallel, without the sequential dependency between time steps required by traditional solvers.

In a context characterized by smooth solutions, uniform spatial coverage, and synthetic data, conventional neural networks demonstrated superior performance, striking a balance between fidelity to the reference solution and computational efficiency. On the other hand, PINNs proved to be more robust in the face of data scarcity, thereby reinforcing their potential for applications in contexts with limited data availability, such as myocardial edema.

Keywords: computational immunology; physics-informed neural networks; computational modeling; infectious myocarditis; myocardial edema.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADAM	*Adaptive Moment Estimation* (Estimativa Adaptativa de Momento)
APC	*Antigen-Presenting Cell* (Célula Apresentadora de Antígeno)
AUC	*Area Under the Curve* (Área Sob a Curva)
CPU	*Central Processing Unit* (Unidade Central de Processamento)
DNN	*Deep Neural Network* (Rede Neural Profunda)
EDO	Equação Diferencial Ordinária
EDP	Equação Diferencial Parcial
FEM	*Finite Element Method* (Método dos Elementos Finitos)
GPU	*Graphics Processing Unit* (Unidade de Processamento Gráfico)
LDNet	*Latent Dynamics Network* (Rede de Dinâmica Latente)
MAE	*Maximum Absolute Error* (Erro Absoluto Máximo)
ML	*Machine Learning* (Aprendizado de Máquina)
MVF	Método dos Volumes Finitos
NEAT	*NeuroEvolution of Augmenting Topologies* (Neuroevolução de Topologias Incrementais)
NN	*Neural Network* (Rede Neural)
NTK	*Neural Tangent Kernel* (Núcleo Tangente Neural)
PAMP	*Pathogen-Associated Molecular Pattern* (Padrão Molecular Associado a Patógenos)
PINN	*Physics-Informed Neural Network* (Rede Neural Informada por Física)
PIGNN	*Physics-Informed Graph Neural Network* (Rede Neural em Grafos Informada por Física)
RMSE	*Root Mean Squared Error* (Raiz do Erro Quadrático Médio)
SGD	*Stochastic Gradient Descent* (Gradiente Descendente Estocástico)

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTO E MOTIVAÇÃO

A miocardite infecciosa é uma condição caracterizada pela inflamação do músculo cardíaco, frequentemente associada à formação de edema miocárdico. A inflamação ocorre quando invasores, também chamados de patógenos, rompem a nossa primeira linha de defesa, composta pelo tecido e pela mucosa, e penetram nos tecidos do organismo. Uma vez no interior do corpo, o patógeno promove sua rápida replicação (no caso de vírus) ou reprodução (no caso de bactérias). No entanto, o sistema imunológico inato está preparado para responder de maneira direta e impedir essa replicação ou reprodução. Em resposta à invasão, células específicas, incluindo os leucócitos, conseguem detectar os patógenos por meio de receptores presentes em suas superfícies. Esses receptores reconhecem substâncias que não estão naturalmente presentes no organismo. Ao detectar a invasão, os leucócitos produzem citocinas pró-inflamatórias para recrutar células adicionais que auxiliam no combate ao patógeno. Essas citocinas desempenham diversas funções; por exemplo, algumas atuam como quimioatraentes, atraindo mais células para o local da infecção. Outras facilitam a movimentação dos leucócitos do sangue, onde estão predominantemente localizados, para o tecido, aumentando a permeabilidade dos vasos sanguíneos. Esse aumento na permeabilidade também permite a saída de fluido derivado do plasma para o espaço intersticial, levando ao acúmulo de líquido intersticial na região e à formação de edema local (LOURENÇO et al., 2022).

O acúmulo excessivo de líquido intersticial pode comprometer a função cardíaca, resultando em arritmias (batimentos cardíacos irregulares), insuficiência cardíaca e até morte súbita. Embora diversos agentes possam desencadear a miocardite, as infecções virais são a causa mais prevalente (CAFORIO et al., 2013). De acordo com a literatura, a incidência de miocardite aumentou em 62,19%, passando de 780.410 casos em 1990 para 1.265.770 casos em 2019, resultando em 324.490 óbitos (WANG et al., 2023). Embora a etiologia da doença seja amplamente compreendida, várias questões fundamentais permanecem sem resposta. O esclarecimento dessas questões poderia melhorar significativamente o cuidado com os pacientes. Nesse contexto, métodos computacionais oferecem ferramentas valiosas para aprofundar a compreensão dos especialistas sobre as interações entre patógenos e o sistema imunológico

Tradicionalmente, a modelagem da interação entre o sistema imunológico e o tecido cardíaco é realizada por meio de equações diferenciais ordinárias (EDOs) e parciais (EDPs), que exigem métodos numéricos computacionalmente intensivos para serem resolvidas com precisão. No entanto, os métodos tradicionais de modelagem e simulação apresentam diversos desafios matemáticos e computacionais. Os valores dos parâmetros, bem como as condições de contorno e iniciais necessárias para a definição completa do modelo,

frequentemente são desconhecidos, o que aumenta a dimensionalidade intrínseca do espaço de soluções. Além disso, o custo computacional associado à aproximação numérica desses modelos pode ser significativo, potencialmente limitando sua aplicabilidade em cenários relevantes (REGAZZONI et al., 2024).

Nos últimos anos, um paradigma antigo conhecido como modelagem baseada em dados tem ganhado força, contrastando com as abordagens tradicionais baseadas na física. Essa mudança foi possibilitada por avanços em otimização, computação de alto desempenho e hardware baseado em GPU, como os núcleos NVIDIA TensorRT, redes neurais artificiais (NNs - *Neural Networks*) e técnicas de Aprendizado de Máquina/Aprendizado Profundo (*Machine/Deep Learning*) (SHAFI et al., 2021). Técnicas baseadas em dados são utilizadas para derivar modelos diretamente a partir de dados experimentais. No entanto, essas abordagens frequentemente carecem de interpretabilidade, tornando difícil a compreensão das relações causais entre os dados observados e as previsões do modelo.

Mesmo diante desses avanços, um dos principais impedimentos à adoção de sistemas baseados em inteligência artificial (IA) é a frequente ausência de transparência. De fato, a natureza de caixa-preta desses sistemas permite realizar previsões poderosas, mas sem possibilidade de explicação direta. Essa limitação desencadeou um novo debate em torno da chamada inteligência artificial explicável (*Explainable AI* – XAI), um campo de pesquisa que demonstra grande potencial para aumentar a confiança e a transparência dos sistemas de IA (ADADI; BERRADA, 2018).

Junto a isso, algumas agências regulatórias passaram, recentemente, a considerar evidências de segurança e eficácia de novos produtos médicos obtidas por meio de modelagem e simulação computacional, prática conhecida como testes *in silico*. Esse movimento despertou o interesse da comunidade de pesquisa em medicina computacional quanto aos aspectos de ciência regulatória dessa disciplina emergente. No entanto, isso impõe um problema fundamental: no domínio da pesquisa biomédica, o uso de modelos computacionais é relativamente recente e ainda carece de um enquadramento epistêmico amplamente aceito para a avaliação da credibilidade dos modelos (VICECONTI et al., 2019).

Nesse contexto, no qual o uso de modelos computacionais usados para testes *in silico* precisam ser confiáveis e auditáveis, os métodos tradicionais baseados em equações dependem de princípios físicos e envolvem a aproximação numérica de equações diferenciais, o que os permite ser altamente interpretáveis, mas exige consideráveis recursos computacionais e conhecimento especializado. Em contraste, os métodos baseados em dados utilizam algoritmos de aprendizado profundo para modelar a dinâmica de sistemas em espaços de menor dimensionalidade. Porém, é mais complexo interpretar as dinâmicas envolvidas nas previsões entregues por modelos baseados em dados. Dessa forma, modelos de aprendizado estatístico enfrentam barreiras na sua implementação em aplicações sensíveis como testes

in silico. Portanto, ao combinar as vantagens de ambas as abordagens, as Redes Neurais Informadas por Física (PINNs - *Physics-Informed Neural Networks*) permitem integrar dados observacionais com restrições físicas, possibilitando a solução de sistemas complexos de equações diferenciais de forma mais eficiente. Diferentemente das abordagens puramente baseadas em dados, que frequentemente carecem de interpretabilidade e demandam grandes quantidades de dados rotulados, as PINNs incorporam conhecimento físico diretamente no treinamento da rede neural, reduzindo a necessidade de dados extensivos e melhorando a generalização do modelo.

Entretanto, apesar do potencial promissor das PINNs em integrar conhecimento físico ao processo de aprendizado, ainda permanecem questões relevantes acerca de sua eficiência prática. A etapa de treinamento dessas redes, ao incorporar derivadas espaciais e temporais das variáveis de interesse, exige o cálculo de gradientes de ordem superior, resultando em um custo computacional significativamente elevado quando comparado ao treinamento de redes neurais convencionais. Ademais, embora as PINNs ofereçam maior interpretabilidade e capacidade de generalização em determinados contextos, ainda não está plenamente estabelecido se seu desempenho, em termos de acurácia e robustez, é de fato superior àquele obtido por abordagens puramente baseadas em dados. Nesse cenário, impõe-se uma reflexão essencial: considerando o esforço computacional envolvido no treino, os métodos de aprendizado profundo estudados realmente oferecem uma vantagem prática significativa sobre métodos tradicionais para modelar a formação de edemas em miocardites infecciosas? Essa indagação constitui a principal motivação do presente trabalho.

1.2 TRABALHOS RELACIONADOS

O modelo utilizado neste estudo deriva de formulações anteriores que descrevem o edema inflamatório com base na teoria da poroelasticidade. Lourenço et al. (2022) propuseram uma formulação baseada no método de elementos finitos com uso de pré-condicionadores eficientes. Outros trabalhos (REIS et al., 2019; REIS et al., 2019; REIS et al., 2018) exploraram modelos similares em uma e duas dimensões espaciais, incorporando a resposta imune. Esses modelos mostraram-se eficazes na simulação do acúmulo de fluido, embora demandem soluções numéricas computacionalmente intensivas e dependam de condições iniciais e de contorno nem sempre disponíveis.

A busca por soluções mais eficientes impulsionou o desenvolvimento de abordagens baseadas em aprendizado de máquina. Neste contexto, destacam-se a proposta de Zohdi (2022), que integra o uso de aprendizado de máquinas (ML - *Machine Learning*) e algoritmos genéticos para simular e otimizar a resposta do sistema imunológico a vacinas. O modelo é resolvido numericamente por meio do Método dos Elementos Finitos (FEM) em malhas bidimensionais e utiliza algoritmos genéticos para otimizar os parâmetros da vacinação, com o objetivo de maximizar a eficácia da resposta imunológica modelada. A principal

inovação reside na integração entre FEM e otimização evolutiva, permitindo simulações rápidas e adaptáveis, tornando o sistema apropriado para a personalização terapêutica. Enquanto o trabalho emprega técnicas de ML para ajustar modelos físicos simplificados, a presente dissertação de mestrado investiga o uso de métodos de aprendizado profundo como alternativa eficiente para formulações mecanicistas baseadas em EDPs. Além disso, os enfoques biológicos diferem: o modelo aqui desenvolvido representa a dinâmica da resposta imune inata, predominante nos estágios iniciais da inflamação, ao passo que o trabalho de Zohdi (2022) concentra-se na resposta adaptativa induzida por vacinas.

Outro exemplo de uso de ML em imunologia é o estudo de Wu et al. (2025), que visa elucidar a participação da necroptose na fisiopatologia da sepse. Utilizando uma abordagem integrativa baseada em transcriptômica e ML, os autores identificaram biomarcadores necroptóticos com potencial diagnóstico. Três genes centrais (*CD40LG*, *TXN* e *AIM2*) foram identificados, com AUC (*Area Under the Curve*) superior a $0,9$ na diferenciação entre pacientes sépticos e controles. Além disso, observou-se forte correlação entre esses genes e alterações no microambiente imunológico, sugerindo sua relevância tanto como marcadores quanto como potenciais alvos terapêuticos. Embora ambos os trabalhos explorem aspectos da resposta inflamatória em contextos infecciosos, a presente dissertação adota uma abordagem mecanicista da dinâmica espaço-temporal da resposta imune, contrastando com a análise molecular e genômica realizada por Wu et al. (2025). Assim, os trabalhos oferecem perspectivas complementares na compreensão das infecções agudas.

No campo do aprendizado profundo, destaca-se a arquitetura *Latent Dynamics Networks* (LDNets) (REGAZZONI et al., 2024), voltada para o aprendizado de dinâmicas espaço-temporais complexas a partir de dados. Tal abordagem permite a construção de modelos de ordem reduzida sem a necessidade de discretizações espaciais explícitas ou codificadores auto-supervisionados. Ao combinar uma rede dinâmica que evolui a solução latente com uma rede de reconstrução espacial consultada ponto a ponto, torna-se possível representar sistemas espaço-temporais de forma contínua e com elevado grau de generalização, mesmo em cenários de extrapolação temporal. As LDNets demonstraram capacidade de superar métodos consagrados de redução de ordem, bem como arquiteturas baseadas em *autoencoders* acoplados a redes dinâmicas, tanto em acurácia quanto em economia de parâmetros. Foram avaliados casos de teste que incluem desde equações diferenciais lineares até modelos altamente não-lineares da eletrofisiologia cardíaca. O trabalho apresenta interseções relevantes com a presente dissertação, sobretudo no propósito de acelerar simulações numéricas e melhorar a representatividade de modelos baseados em equações diferenciais por meio de técnicas de aprendizado profundo. Ambas as abordagens buscam explorar representações latentes compactas que substituam métodos numéricos tradicionais, promovendo ganhos computacionais expressivos. Contudo, distinguem-se em diversos aspectos fundamentais. Enquanto a LDNets são treinadas diretamente sobre

dados espaço-temporais e realizam inferência em pontos arbitrários do domínio por meio de uma arquitetura *meshless*, a presente dissertação adota a arquitetura *meshless* no cálculo dos resíduos das EDPs; para a perda de dados adota-se uma formulação explícita com discretização fixa do domínio na perda de dados. Nesse contexto, o treinamento *meshless* aumenta a capacidade de generalização do modelo, uma vez que este é apresentado a pontos que não pertencem à malha adotada nos modelos discretos, e pode reduzir o custo computacional do treino, visto que podemos usar menos pontos por iteração. Além disso, LDNets incorporam um modelo dinâmico explícito no espaço latente, o que não se observa na arquitetura PINN adotada nesta dissertação, onde a dinâmica do sistema é imposta por penalização dos resíduos das EDPs na função de perda. Ambas as abordagens reforçam a tendência emergente de integrar conhecimento físico e métodos de aprendizado de máquina para modelar sistemas dinâmicos complexos com menor custo computacional, ainda que com estratégias distintas de representação e inferência das soluções.

No que diz respeito a técnicas de treinamento de redes neurais, o algoritmo NEAT (NeuroEvolution of Augmenting Topologies) (MIIKKULAINEN et al., 2019) é uma abordagem evolucionista que combina a otimização dos pesos com a evolução progressiva da topologia da rede. Diferentemente dos métodos tradicionais, nos quais a arquitetura é fixada *a priori*, o NEAT inicia o processo evolutivo com redes de estrutura simples, que são gradualmente complexificadas por meio da inserção de novos nós e conexões, conforme a pressão seletiva favorece topologias mais eficazes. O algoritmo se apoia na codificação genética com rastreamento de inovações, que assegura a viabilidade do cruzamento entre redes de diferentes topologias, na especiação, que preserva estruturas emergentes promissoras ao evitar sua eliminação prematura e na complexificação incremental, que promove o crescimento controlado da rede sem comprometer a estabilidade do processo evolutivo. Os autores demonstram a eficácia dessa estratégia em tarefas dinâmicas não triviais, como o controle de sistemas parcialmente observáveis, superando abordagens clássicas de *neuroevolution* em diversos cenários. Embora o presente trabalho adote uma abordagem, baseada em arquiteturas fixas e treinamento por gradiente utilizando o otimizador ADAM (*Adaptive Moment Estimation* — Estimativa adaptativa de momento), há convergência entre os objetivos metodológicos de ambas as propostas. Em particular, a motivação para utilizar redes capazes de representar fenômenos complexos e não lineares de forma eficiente é compartilhada. No entanto, ao contrário do NEAT, que se destaca por adaptar dinamicamente sua estrutura topológica em resposta ao desempenho evolutivo, a abordagem aqui empregada assume uma arquitetura previamente definida. Além disso, o contexto de aplicação é substancialmente diferente, enquanto o NEAT é orientado a problemas de controle e tomada de decisão em ambientes com representação parcial do estado, o presente estudo se insere no escopo da modelagem de fenômenos físico-biológicos governados por EDPs, nos quais a consistência com as leis diferenciais é fundamental. Nesse cenário, a capacidade do NEAT de otimizar simultaneamente pesos e topologia

aponta para um caminho promissor em aplicações futuras que busquem automatizar a escolha da arquitetura mais adequada.

Dentre as abordagens de ML, as PINNs têm se destacado por permitir a incorporação de leis físicas diretamente no processo de treinamento da rede neural. Essa abordagem híbrida oferece uma alternativa promissora à modelagem puramente empírica ou tradicional, sendo capaz de lidar com dados escassos e fornecer maior interpretabilidade ao modelo. Tal interpretabilidade advém do fato de que a diferenciação automática utilizada no treinamento da rede deve respeitar, obrigatoriamente, as equações diferenciais que regem o fenômeno modelado. Diferentemente de redes puramente estatísticas, os pesos das PINNs estão intrinsecamente vinculada à estrutura matemática do sistema físico, o que permite, por exemplo, a aplicação da chamada PINN inversa, na qual parâmetros desconhecidos das equações são inferidos diretamente a partir de dados observacionais. Esse alinhamento explícito entre a rede e a modelagem mecanicista favorece tanto a interpretabilidade quanto a confiabilidade do modelo gerado, aproximando-o de aplicações científicas e clínicas sensíveis. Ainda, PINNs demonstraram ser até seis vezes mais rápidas que algoritmos baseados em métodos tradicionais de resolução numérica, como diferenças finitas, ao resolver um modelo espaço-temporal da resposta inflamatória (FERNANDES et al., 2024). Outra de suas aplicações é a modelagem e otimização do processo de cura fora de autoclave de compósitos termofixos (HUMFELD et al., 2025). A formulação também explora novas formas de melhorar o treinamento e emprega múltiplas redes neurais co-treinadas para representar, de forma autônoma, as variáveis físicas do sistema, como temperatura do ar, da peça, da ferramenta e grau de cura, viabilizando a imposição de restrições tanto sobre entradas quanto sobre saídas. A estrutura diferencial do problema envolve equações de transferência de calor com reação exotérmica e condições de contorno do tipo Robin, resolvidas de modo acoplado. O processo de treinamento ocorre em modos sucessivos, nos quais as equações físicas e restrições operacionais são introduzidas progressivamente, garantindo estabilidade e convergência. O perfil de temperatura do ar otimizado respeita os limites impostos de temperatura, taxa de aquecimento/resfriamento e grau de cura, resultando em tempo de cura reduzido e boa aderência às soluções obtidas por FEM e por experimentos com placas compostas. Erros médios de 5,5% na previsão da temperatura e 2,2% no grau de cura foram observados. A formulação permite eliminar a necessidade de buscas iterativas, típicas de abordagens tradicionais, ao produzir soluções compatíveis com os dados experimentais e com as equações governantes de forma unificada. Contudo, enquanto o artigo propõe uma abordagem para otimização simultânea de entradas e saídas físicas por meio do co-treinamento de múltiplas redes, ela difere da presente dissertação ao não realizar comparação sistemática com outras abordagens (como MVF e NNs) nem abordar explicitamente os desafios da escassez de dados em modelos espaço-temporais da resposta imunológica.

Também foi proposta na literatura uma arquitetura baseada em PINNs aplicada

à modelagem da eletrofisiologia cardíaca (MARTIN et al., 2022). Essa abordagem visa inferir propriedades eletrofisiológicas do tecido, como excitabilidade, coeficientes de difusão e duração do potencial de ação, a partir de medidas escassas do potencial transmembrânico. Utilizando o modelo de Aliev-Panfilov em geometrias unidimensionais e bidimensionais, o método foi avaliado em contextos homogêneos, heterogêneos e sob condições arrítmicas (espirais), sendo capaz de reproduzir com boa acurácia os padrões espaço-temporais da dinâmica elétrica cardíaca. Os testes foram realizados tanto com dados sintéticos quanto com dados experimentais adquiridos por mapeamento óptico, evidenciando a capacidade das PINNs de estimar parâmetros de modelos em situações reais, inclusive em resposta a fármacos antiarrítmicos. A formulação demonstra robustez frente à escassez de dados e ruídos experimentais, com desempenho superior ao de métodos inversos tradicionais em termos de generalização e custo computacional. Embora haja semelhança metodológica com o presente trabalho quanto ao uso de PINNs para modelagem baseada em EDPs, os objetivos e sistemas estudados são distintos: o estudo em questão visa a eletrofisiologia do miocárdio, enquanto a presente dissertação foca na dinâmica espaço-temporal da resposta imunológica inata. Apesar das diferenças nos sistemas modelados, a estratégia adotada nesse estudo fornece uma motivação relevante para o presente trabalho, ao demonstrar que PINNs podem ser aplicadas com sucesso à inferência de parâmetros em sistemas fisiológicos complexos e ruidosos. A capacidade de lidar com dados escassos e reproduzir padrões espaço-temporais a partir de observações limitadas inspira a aplicação de técnicas semelhantes na modelagem da resposta imune inata, especialmente em cenários onde a obtenção de dados experimentais é restrita ou invasiva.

Outro trabalho propõe um modelo compartimental para descrever a dinâmica epidemiológica da variante Ômicron da COVID-19 e suas sublinhagens (ZOHDI, 2022). A formulação matemática incorpora compartimentos clássicos (Suscetíveis, Infectados, Hospitalizados, Recuperados, Óbitos e Vacinados), cujas transições são regidas por um sistema de equações diferenciais ordinárias. A arquitetura PINN é treinada para resolver o problema inverso de inferência paramétrica, estimando dinamicamente, a cada janela deslizante de 90 dias, variáveis como taxa de transmissão, reinfecção e eficácia vacinal com base em dados públicos de três países europeus. Para isso, a função de perda combina termos de erro relativos aos dados observados e aos resíduos das equações diferenciais, sendo seu balanceamento controlado automaticamente por um parâmetro adaptativo que ajusta, a cada época, a razão entre as normas dos gradientes. Os resultados evidenciam a capacidade do método em capturar a evolução dos parâmetros durante a circulação de diferentes sublinhagens da Ômicron, com destaque para a mutação S371F, associada ao aumento da taxa de transmissão. Ademais, demonstrou-se o potencial do modelo como ferramenta preditiva voltada ao apoio à tomada de decisão em saúde pública. Apesar de ambos os estudos explorarem o uso de PINNs para modelar processos biológicos mediados por equações diferenciais, há distinções significativas, entre o artigo em questão e a

presente dissertação, em termos de escopo, objetivos e escala do fenômeno investigado. Enquanto o artigo analisa a progressão macroscópica da pandemia por meio de séries temporais populacionais, este trabalho foca na modelagem espaço-temporal da resposta imunológica inata a infecções virais, com ênfase na dinâmica de patógenos e leucócitos no tecido miocárdico. Em termos de métodos, a presente dissertação também realiza uma comparação entre PINNs, Redes Neurais convencionais e o MVF, considerando aspectos como o custo computacional e fidelidade da solução de referência, o que não é abordado no estudo da Ômicron.

Por fim, destaca-se a arquitetura *Physics-Informed Graph Neural Network* (PIGNN), proposta por Chen et al. (2025) para simular escoamentos monofásicos em meios porosos com geometrias irregulares. O modelo combina a estrutura das redes em grafos com os princípios das PINNs, utilizando esquemas numéricos explícitos (MVF, média harmônica, diferenciação a montante) incorporados em camadas convolucionais para representar com precisão o fluxo entre células adjacentes. A PIGNN demonstrou desempenho superior às PINNs tradicionais em termos de erro relativo, estabilidade temporal e capacidade de generalização. A convergência conceitual com esta dissertação é notável: ambos os trabalhos modelam sistemas físicos em meios porosos, com preocupação explícita quanto à fidelidade local da solução, robustez frente à escassez de dados e equilíbrio entre custo computacional e acurácia. Ainda que a presente dissertação não aborde diretamente os aspectos mecânicos da formação de edemas, os resultados da PIGNN indicam o potencial de expansão do modelo atual para representar de forma unificada os diversos aspectos do processo inflamatório.

Embora a literatura recente evidencie avanços significativos na aplicação de técnicas de aprendizado de máquina à modelagem de processos biológicos, constata-se que o uso de PINNs permanece relativamente pouco apresentado no contexto da modelagem de respostas imunológicas, principalmente do sistema imune inato. A maior parte dos estudos concentra-se em abordagens convencionais baseadas em dados ou em estratégias híbridas voltadas à calibração de modelos físicos simplificados, sem, contudo, explorar comparações sistemáticas entre PINNs e outros métodos de resolução, como NNs e MVF. Esta lacuna limita a compreensão crítica acerca da efetividade prática das redes neurais profundas (DNNs - *Deep Neural Networks*), especialmente em termos de custo computacional, fidelidade local da solução e robustez frente à escassez de dados. Nesse sentido, a presente dissertação insere-se de forma original e relevante ao propor uma análise comparativa entre essas abordagens no contexto da modelagem espaço-temporal da resposta imunológica inata, contribuindo para elucidar as condições sob as quais a adoção de aprendizado profundo se mostra vantajoso frente a métodos tradicionais.

1.3 OBJETIVOS

O principal objetivo deste trabalho é investigar a aplicação de DNNs na modelagem da dinâmica da resposta imune local em casos de miocardite infecciosa. Especificamente, busca-se representar as concentrações de patógenos e leucócitos ao longo do tempo e do espaço por meio de uma formulação baseada em PINNs e NNs. Como objetivos complementares, pretende-se comparar o desempenho dessas abordagens com o MVF, além de avaliar como a quantidade de amostras de treinamento e a arquitetura da rede influenciam a acurácia e o custo computacional das soluções obtidas.

1.4 PRINCIPAIS CONTRIBUIÇÕES

As contribuições deste trabalho concentram-se em quatro frentes principais. Primeiramente, desenvolveu-se e implementou-se uma arquitetura baseada em PINNs para modelar a dinâmica espaço-temporal da resposta imunológica em tecidos miocárdicos, incorporando explicitamente as equações diferenciais que regem a evolução de patógenos e leucócitos. Em seguida, emprega-se um protocolo comparativo sistemático entre abordagens baseadas em PINNs e NNs, considerando critérios como acurácia, robustez frente à variação da densidade amostral e custo computacional de treinamento, com o intuito de promover uma avaliação crítica da aplicabilidade de PINNs em contextos de imunologia computacional.

Adicionalmente, este trabalho apresenta evidências empíricas de que as PINNs são menos sensíveis à redução da densidade amostral quando comparadas às NNs, embora apresentem limitações na representação de dinâmicas locais e estruturas estacionárias. Por fim, demonstra-se que, em problemas caracterizados por soluções suaves e boa cobertura amostral, redes neurais tradicionais podem superar as PINNs em termos de desempenho, contrariando a expectativa frequentemente atribuída à superioridade generalizada das redes informadas por física (FERNANDES et al., 2024).

1.5 ORGANIZAÇÃO DO TEXTO

Este trabalho está estruturado de maneira a conduzir o leitor progressivamente desde os fundamentos conceituais até os resultados obtidos por meio da implementação computacional de uma versão simplificada do modelo utilizado para descrever a resposta imune a um patógeno com o emprego de NNs e PINNs, promovendo uma compreensão clara e contextualizada das contribuições apresentadas.

No presente Capítulo 1, apresenta-se a motivação para o estudo, evidenciando a relevância da modelagem matemática e computacional da resposta imunológica. São descritos os objetivos da pesquisa, bem como suas principais contribuições no contexto da aplicação de métodos baseados em redes neurais, com destaque para as PINNs.

O Capítulo 2 é dedicado ao referencial teórico necessário para a compreensão do trabalho. Inicialmente, descreve-se o funcionamento do sistema imune inato e os mecanismos biológicos envolvidos na resposta inflamatória. Complementarmente, apresenta-se o embasamento teórico necessário para o entendimento das redes neurais artificiais, do processo de retropropagação do erro, dos métodos de otimização utilizados e, por fim, a formulação das PINNs.

O método adotado na implementação computacional, bem como as rotinas desenvolvidas para avaliação do desempenho dos modelos, é descrito no Capítulo 3. Neste capítulo apresenta-se o modelo matemático utilizado para representar a dinâmica espaço-temporal da concentração de patógenos e leucócitos no tecido intersticial. São também discutidas as suposições adotadas, a discretização das equações via o MVF e as condições de estabilidade numérica. O Capítulo 4 apresenta os resultados obtidos, organizados em duas seções principais: a primeira dedicada à busca em grade de hiperparâmetros para as arquiteturas PINN, e a segunda à comparação entre os modelos considerados (MVF, NN e PINN), avaliando-se a capacidade de representação, a sensibilidade à redução de amostras temporais e o desempenho computacional.

Finalmente, o Capítulo 5 reúne as considerações finais do trabalho, destacando as conclusões extraídas a partir dos experimentos realizados, as limitações identificadas e as perspectivas para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos necessários para o entendimento desta dissertação de mestrado. Inicialmente apresentaremos uma breve revisão do funcionamento do sistema imunológico humano. Na sequência abordaremos o funcionamento do aprendizado supervisionado, das redes neurais e a sua variação, redes neurais informadas por física.

2.1 SISTEMA IMUNOLÓGICO

A proteção do organismo frente a infecções depende de uma complexa rede de defesas, que atua para impedir a entrada e proliferação de agentes patogênicos. Em geral, infecções se iniciam quando vírus, bactérias ou outros microrganismos patogênicos conseguem ultrapassar as barreiras do corpo humano e estabelecem um foco de replicação ou reprodução (SOMPAYRAC, 2010). A entrada desses patógenos pode ocorrer por múltiplas vias, incluindo o contato direto com fluidos corporais, lesões cutâneas, uso de instrumentos médicos contaminados, ou ainda através de superfícies mucosas expostas, como olhos e boca.

2.1.1 Sistema imune inato

Como primeira linha de defesa, o organismo conta com barreiras físicas e químicas, como a pele, as mucosas, proteínas solúveis e células especializadas, que em conjunto constituem o sistema imune inato.

A pele humana, com área estimada em cerca de 2 m^2, funciona como uma barreira mecânica eficaz contra a maioria dos microrganismos (SOMPAYRAC, 2010). As mucosas, por sua vez, revestem os tratos digestivo, respiratório e reprodutivo, totalizando uma área consideravelmente maior, de aproximadamente 400 m^2, e apresentam características químicas que dificultam a sobrevivência de microrganismos invasores, como a acidez do ambiente e a secreção de muco.

Apesar de sua relevância, essas barreiras não são infalíveis. Diversos patógenos desenvolveram mecanismos que lhes permitem evadir essas defesas iniciais. Quando tal evasão ocorre, entram em ação proteínas do sistema complemento e células especializadas do sistema imune inato, que respondem de forma rápida ao reconhecer padrões moleculares associados a patógenos (SOMPAYRAC, 2010). Modelos computacionais têm sido utilizados para investigar a dinâmica dessa resposta (FERNANDES et al., 2024; LOURENÇO et al., 2022), sendo os de especial interesse deste trabalho aqueles focados em contextos inflamatórios agudos, como na formação de edemas, contribuindo para o entendimento das interações entre agentes infecciosos e células do sistema imunológico (FERNANDES et al., 2024).

O sistema imune inato constitui, portanto, a primeira barreira biológica efetiva contra microrganismos invasores, desencadeando respostas imediatas logo após o reconhecimento de padrões moleculares conservados característicos de patógenos (SOMPAYRAC, 2010). Diversas populações de leucócitos, células originadas na medula óssea a partir de células-tronco hematopoéticas, compõem esse sistema.

Entre as suas principais funcionalidades destaca-se a fagocitose, processo pelo qual células fagocitárias, como neutrófilos, monócitos/macrófagos, mastócitos e células dendríticas, internalizam partículas estranhas em vesículas denominadas fagossomos. Posteriormente, os fagossomos fusionam-se a lisossomos, que são compartimentos ricos em enzimas hidrolíticas responsáveis pela degradação do patógeno. Além de eliminarem microrganismos, muitos fagócitos atuam como células apresentadoras de antígeno (APCs), expondo peptídeos derivados do invasor a células da imunidade adaptativa, como os linfócitos T e, dessa forma, estabelecendo o elo funcional entre imunidade inata e adaptativa.

Embora o sistema inato seja extremamente eficiente frente à maioria dos patógenos, certos agentes, notadamente vírus ou bactérias intracelulares, podem escapar a essa linha de defesa. Nesses cenários, ativa-se o sistema adaptativo, capaz de gerar respostas altamente específicas e de longa duração. Este trabalho focará no sistema inato, motivo pelo qual não apresentaremos os detalhes do funcionamento do sistema adaptativo. O leitor mais interessado poderá encontrar detalhes deste assunto em livros da área de imunologia (SOMPAYRAC, 2010).

2.1.2 Resposta inflamatória

A transposição da barreira físico-química por microrganismos desencadeia uma sequência coordenada de eventos conhecida como resposta inflamatória. O processo é deflagrado quando células sentinelas presentes nos tecidos, em especial os macrófagos, reconhecem estruturas moleculares estranhas (PAMPs - *Pathogen-Associated Molecular Patterns*) por meio de receptores de reconhecimento padrão em sua superfície (SOMPAYRAC, 2010). Uma vez ativados, esses fagócitos liberam citocinas e quimiocinas pró-inflamatórias que promovem o recrutamento de leucócitos adicionais, induzem vasodilatação e aumentam a permeabilidade do endotélio.

O extravasamento de plasma e células de defesa para o tecido lesado leva a acumulação de fluido intersticial, resultando na formação de um edema. Simultaneamente, o maior fluxo sanguíneo causa rubor e elevação local da temperatura, enquanto mediadores químicos ativam terminações nervosas provocando dor. Assim, edema, vermelhidão, calor e dor compõem a tétrade clássica da inflamação aguda (SOMPAYRAC, 2010).

À medida que o agente invasor é neutralizado, a mesma rede de citocinas passa a induzir a secreção de mediadores anti-inflammatórios, como IL-10, TGF-β e lipoxinas, que restauram a integridade vascular e sinalizam o término do recrutamento celular.

Sob esse novo microambiente, os leucócitos efetores, em especial neutrófilos e alguns monócitos recém-chegados, entram em apoptose e exibem fosfatidilserina na face externa da membrana, o que os marca para remoção. Os macrófagos residentes ou derivados de monócitos reconhecem esses sinais por receptores específicos e realizam a eferocitose, fagocitando as células apoptóticas e detritos teciduais.

Esse processo, além de prevenir a liberação de conteúdo citotóxico, reprograma funcionalmente os próprios macrófagos, que passam a produzir mediadores pró-resolução – resolvinas, maresinas e protectinas – promovendo a cessação da resposta inflamatória. O excesso de fluido intersticial é então drenado pela vasculatura linfática, o número de células imunes retorna a níveis basais e consolida-se a memória imunológica do episódio, permitindo respostas mais rápidas em exposições subsequentes. Desse modo, por meio da orquestração entre apoptose, eferocitose e drenagem linfática, o sistema imune conclui a resposta inflamatória e restabelece a homeostase tecidual.

2.2 APRENDIZADO SUPERVISIONADO

O aprendizado supervisionado constitui uma das abordagens fundamentais em aprendizado de máquina, caracterizando-se pelo uso de dados rotulados para treinar modelos preditivos. Nesse contexto, cada entrada do conjunto de dados é associada a uma saída esperada, permitindo ao modelo aprender uma função de mapeamento capaz de generalizar para amostras não vistas. Formalmente, seja $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ o conjunto de treinamento, onde $\mathbf{x}_i \in \mathbb{R}^m$ representa um vetor de atributos com m dimensões, $\mathbf{y}_i \in \mathbb{R}^j$ é o vetor de saída correspondente com j dimensões, e n é o número total de amostras. O objetivo do treinamento supervisionado é encontrar uma função $f_{\boldsymbol{\theta}}$, parametrizada por $\boldsymbol{\theta} \in \mathbb{R}^p$ que minimize o erro entre as saídas preditas $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ e os rótulos reais $\mathbf{y}_i$. Essa discrepância é quantificada por uma função de perda $\mathcal{L}(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\theta}_t)$, onde o índice t denota a iteração atual do processo de otimização (GOODFELLOW et al., 2016).

A Figura 1 apresenta uma visão esquemática do processo de aprendizado supervisionado adotado em diversos métodos de aprendizagem de máquina. Inicialmente, os dados são extraídos de uma base e particionados em subconjuntos de treinamento e validação. Cada amostra é representada por um vetor de atributos, acompanhado de um rótulo correspondente. Durante o treinamento, apenas os dados de treinamento são utilizados para ajustar iterativamente os parâmetros da rede por meio de um algoritmo de otimização, até que um critério de parada seja atingido, como a convergência da função de perda ou o número máximo de épocas. Ao término do processo, o modelo treinado é avaliado utilizando-se o conjunto de validação, com o intuito de aferir sua capacidade de generalização e a fidelidade da função aprendida frente a dados não vistos. Uma vez validado, o modelo pode ser empregado para realizar inferências sobre novas entradas.

Esse paradigma é amplamente empregado em tarefas de classificação e regressão. No

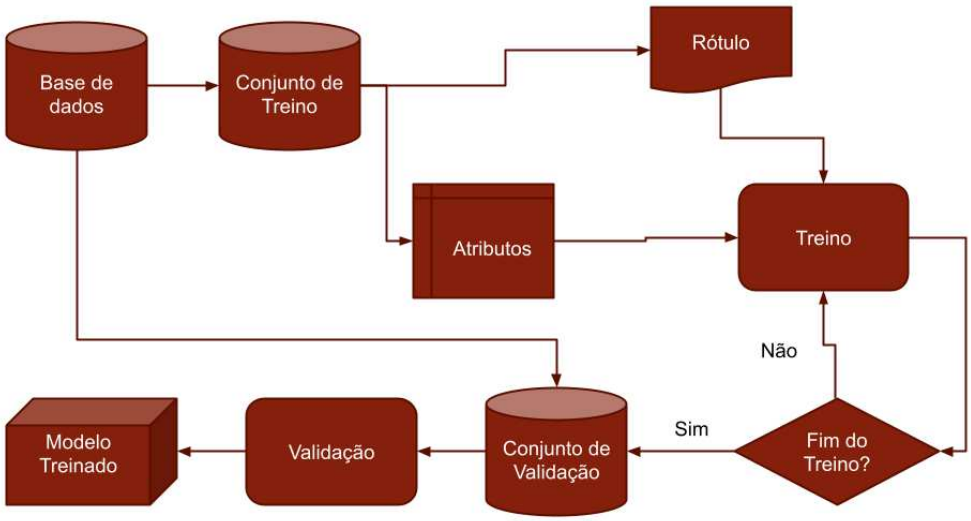


Figura 1 – Fluxo do processo de aprendizado supervisionado.

contexto do presente trabalho, redes neurais supervisionadas são treinadas para aproximar as soluções de um sistema dinâmico governado por equações diferenciais parciais, a partir de dados sintéticos gerados a partir da resolução destas equações usando o MVF.

2.3 REDES NEURAIS E O PROCESSO DE TREINAMENTO

Redes neurais artificiais são modelos computacionais inspirados na estrutura e no funcionamento dos neurônios biológicos. Em termos abstratos, cada neurônio artificial recebe múltiplas entradas ponderadas, aplica uma transformação linear seguida de uma não linearidade, e propaga o resultado para outras unidades. A Figura 2 estabelece uma analogia entre o neurônio biológico e sua contraparte computacional: os dendritos, responsáveis por captar sinais de outras células, correspondem às entradas $x_1, x_2, \ldots, x_n$ multiplicadas por seus respectivos pesos sinápticos $w_1, w_2, \ldots, w_n$; o corpo celular, que realiza a soma dos sinais recebidos, é análogo ao somatório linear $\sum w_j x_j + b$, com b representando o viés; a bainha de mielina, que modula a propagação do impulso elétrico ao longo do axônio, tem papel similar ao da função de ativação $\psi(\cdot)$, que determina a intensidade e continuidade da propagação da saída. Por fim, o axônio equivale à saída do neurônio, que se conecta a outros neurônios da camada seguinte.

O objetivo de uma rede neural é aprender uma função de mapeamento entre entradas e saídas, ajustando iterativamente os pesos $\boldsymbol{w}$ e o viés b de cada neurônio, de forma a minimizar o erro entre a saída estimada e a saída desejada para um conjunto de exemplos observados. Esse processo de ajuste é realizado por meio de algoritmos de otimização baseados em gradiente, como o método do gradiente descendente (GOODFELLOW et al., 2016).

Redes neurais artificiais do tipo *feedforward*, como ilustrado na Figura 3, são

compostas por uma sequência de camadas organizadas de forma hierárquica. A camada de entrada é responsável por receber os atributos de cada amostra, representados por um vetor $\mathbf{x}_i \in \mathbb{R}^m$. Esses valores são propagados para uma ou mais camadas ocultas, cujos neurônios realizam transformações lineares seguidas da aplicação de uma função de ativação não linear $\psi(\cdot)$. Finalmente, a camada de saída fornece a estimativa $\hat{\mathbf{y}}_i \in \mathbb{R}^j$, correspondente à predição do modelo para aquela entrada.

O número de camadas ocultas (profundidade) e a quantidade de neurônios por camada (largura) influenciam diretamente a capacidade expressiva da rede. Arquiteturas mais profundas são capazes de representar composições funcionais mais complexas, enquanto redes mais largas ampliam o espaço de funções acessível ao modelo. No entanto, ambos os casos demandam estratégias de regularização adequadas para mitigar riscos de sobreajuste (GOODFELLOW et al., 2016).

Matematicamente, uma rede neural com L camadas pode ser representada como uma composição de funções:

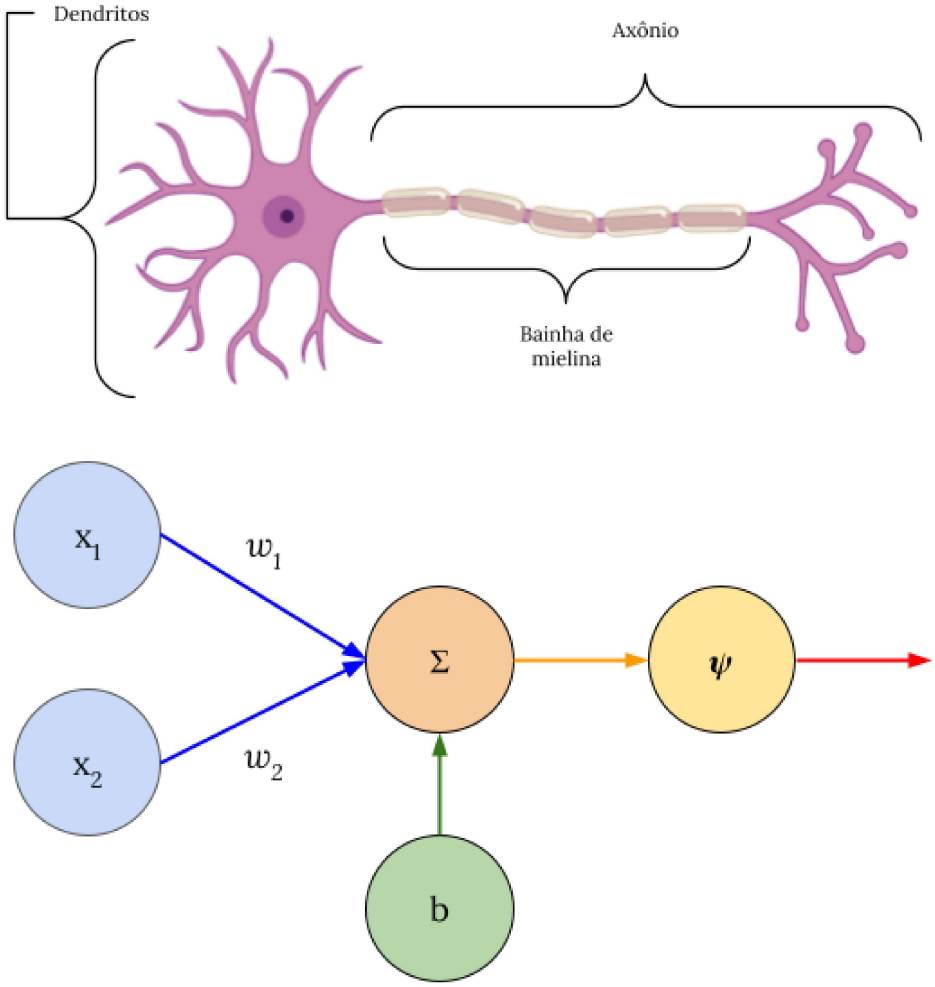


Figura 2 – Analogia entre o neurônio biológico e o neurônio artificial.

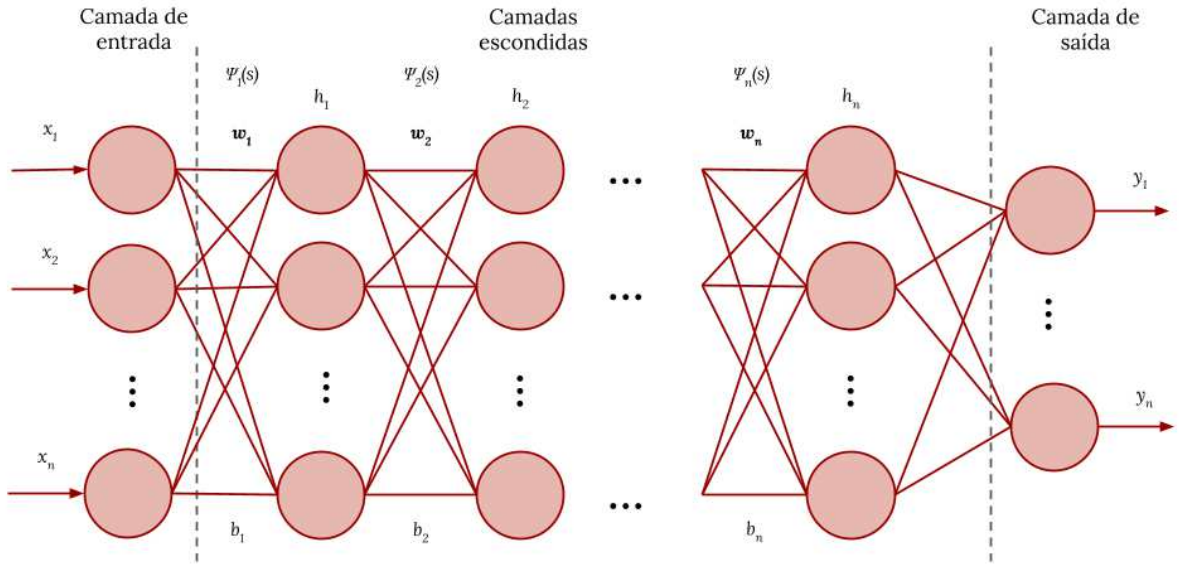


Figura 3 – Arquitetura típica de uma rede neural artificial do tipo *feedforward*.

$$f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \mathbf{h}^{(L)} = \psi^{(L)}\left(\psi^{(L-1)}\left(\cdots \psi^{(1)}(\mathbf{x}_i)\cdots\right)\right), \tag{2.1}$$

onde cada transformação $\phi^{(l)}$ é dada por:

$$\phi^{(l)}(\mathbf{h}^{(l-1)}) = \psi^{(l)}\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right), \tag{2.2}$$

sendo $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ e $\mathbf{b}^{(l)} \in \mathbb{R}^{d_l}$ os pesos e vieses da l-ésima camada, $\psi^{(l)}(\cdot)$ a função de ativação, e $\boldsymbol{\theta} = \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}_{l=1}^{L}$ o conjunto de parâmetros treináveis. Dessa forma, a rede realiza sucessivas transformações não lineares até produzir a predição final.

Uma vez definido o processo de inferência $f_\theta(\mathbf{x_i})$, passa-se para o processo de aprendizado. O processo de aprendizado supervisionado mais utilizado nas redes neurais *feedforward* é baseado no método do gradiente descendente estocástico (SGD – *Stochastic Gradient Descent*), aliado ao algoritmo de retropropagação do erro (*backpropagation*). A ideia central do *backpropagation* é propagar o erro da saída para as camadas anteriores da rede, possibilitando a atualização dos pesos de forma eficiente por meio do cálculo do gradiente da função de perda em relação a esses pesos.

Formalmente, seja $\boldsymbol{x} \in \mathbb{R}^n$ um vetor de entrada, $\boldsymbol{y} \in \mathbb{R}^m$ a saída desejada e $f(\boldsymbol{x}; \boldsymbol{\theta})$ a saída produzida por uma rede com parâmetros $\boldsymbol{\theta}$ (pesos e vieses). Define-se uma função de perda $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ que quantifica o erro de predição. A cada iteração t, um novo par $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ é apresentado à rede, e os parâmetros são atualizados segundo a regra:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \mathbf{C} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\theta}_t), \tag{2.3}$$

onde α_t é a taxa de aprendizado e $\mathbf{C}$ é uma matriz simétrica definida positiva. Essa matriz é chamada de pré-condicionador pois atua modificando a direção e a escala do gradiente, adaptando o passo de atualização à geometria local da função de perda. Em outras palavras, $\mathbf{C}$ transforma o gradiente para compensar a má-condição do problema de otimização — por exemplo, quando a superfície de perda é alongada em certas direções — permitindo uma convergência mais rápida. No caso particular em que $\mathbf{C} = \mathbf{I}_d$, onde $\mathbf{I}_d$ denota a matriz identidade de ordem d, obtém-se a forma tradicional do método de gradiente descendente estocástico, sem qualquer ajuste direcional (AMARI, 1993).

Essa formulação geral permite compreender o método de retropropagação como um caso particular de um algoritmo mais amplo de otimização estocástica aplicado a redes parametrizadas. A ampla aplicabilidade do algoritmo de retropropagação a diversos modelos de rede e funções de perda tornou-o um pilar fundamental no treinamento de redes neurais. De fato, versões iniciais desse método já haviam sido estudadas na década de 1960 no contexto de redes multicamadas com aprendizado baseado em gradiente (AMARI, 1993).

Contudo, o método de gradiente descendente simples apresenta limitações práticas, como sensibilidade à escala dos gradientes, convergência lenta e dificuldades em escapar de platôs ou mínimos locais rasos. Para mitigar essas limitações, algoritmos mais sofisticados utilizam versões adaptativas ou aproximadas de $\mathbf{C}$ com o objetivo de ajustar dinamicamente o aprendizado às curvaturas do espaço de parâmetros. Entre eles, destaca-se o algoritmo ADAM, que combina estimativas adaptativas de momentos de primeira e segunda ordem do gradiente.

Adicionalmente, o termo "*estocástico*" no contexto do SGD refere-se à prática de calcular os gradientes com base em *batchs* aleatórios de dados, em vez de utilizar todo o conjunto de treinamento a cada iteração. Essa amostragem introduz variabilidade no processo de otimização, o que pode favorecer a generalização e auxiliar na fuga de mínimos locais rasos (BOTTOU, 2010).

2.4 ESTIMATIVA ADAPTATIVA DE MOMENTO

O algoritmo ADAM é um método de otimização estocástica de primeira ordem amplamente utilizado no treinamento de DNNs. Ele foi proposto por Kingma e Ba (2015) como uma alternativa robusta e eficiente para problemas com grandes volumes de dados, alta dimensionalidade e gradientes ruidosos ou esparsos. Diferentemente do gradiente descendente clássico, que utiliza uma única taxa de aprendizado fixa para todos os parâmetros, o ADAM adota uma estratégia adaptativa que ajusta dinamicamente os passos de atualização com base em estimativas dos momentos de primeira (média) e segunda ordem (variância) dos gradientes (DUCHI; HAZAN; SINGER, 2011; GOODFELLOW et al., 2016).

Essa estimativa é realizada separadamente para cada parâmetro e varia ao longo das iterações, permitindo que o algoritmo se adapte às características locais da superfície de erro. No contexto da formulação geral apresentada na Equação 2.3, essa abordagem equivale à construção implícita de uma matriz pré-condicionadora $\mathbf{C}$, cujos coeficientes são calculados a partir das estatísticas acumuladas dos gradientes.

Tal precondicionamento acelera a convergência, especialmente em domínios com topologia complexa. Isso é particularmente relevante para o treinamento de PINNs, cuja função de perda é composta por múltiplos termos com escalas distintas, incluindo resíduos de equações diferenciais, condições de contorno, condições iniciais e dados observacionais, o que gera um problema de otimização altamente complexo. Nesses casos, métodos com atualização adaptativa, como o ADAM, contribuem para estabilizar o processo de treinamento e mitigar a dominância de termos específicos da perda.

O algoritmo mantém estimativas exponencialmente decrescentes das médias dos gradientes (m_t) e dos quadrados dos gradientes (v_t), conforme descrito pelas equações:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \tag{2.4}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \tag{2.5}$$

onde g_t representa o gradiente da função objetivo no tempo t, ou $\nabla_{\boldsymbol{\theta}} \boldsymbol{\mathcal{L}}(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\theta}_t)$ na Equação 2.3, e os hiperparâmetros β_1 e β_2 controlam a taxa de decaimento das médias móveis. Esses valores são tipicamente próximos de 1 (por exemplo, $\beta_1 = 0{,}9$ e $\beta_2 = 0{,}999$), conforme sugerido por Kingma e Ba (2015). Neste trabalho, tais hiperparâmetros farão parte de uma busca em grade, abordada na Seção 4.2, com o intuito de determinar a melhor combinação para o problema proposto.

Devido à inicialização em zero, as estimativas de m_t e v_t são tendenciosas nos primeiros passos do treinamento. Para corrigir esse viés, são computadas as versões corrigidas:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, e \tag{2.6}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}. \tag{2.7}$$

A atualização dos parâmetros do modelo é então realizada segundo a fórmula:

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}, \tag{2.8}$$

em que α é a taxa de aprendizado e ε é um pequeno valor positivo adicionado para garantir estabilidade numérica.

O ADAM apresenta desempenho particularmente eficaz em problemas com gradientes ruidosos ou esparsos, adaptando automaticamente as taxas de aprendizado individuais para cada parâmetro. Essa característica favorece uma convergência mais rápida e estável, sendo especialmente valiosa em aplicações com redes profundas e estruturas paramétricas complexas. Ainda segundo Kingma e Ba (2015), o método mostra vantagens em relação a otimizadores convencionais mesmo em contextos não estacionários.

2.5 REDES NEURAIS INFORMADAS POR FÍSICA (PINN)

Esta seção aborda a resolução de EDPs por meio da abordagem conhecida como PINNs. Uma PINN é um tipo especializado de rede neural que incorpora diretamente as leis da física no processo de aprendizado. Diferentemente das redes neurais tradicionais, que dependem exclusivamente de dados observados para ajustar seus parâmetros, as PINNs utilizam o conhecimento físico do sistema como parte da função de perda. Essa estratégia garante que a solução obtida esteja não apenas ajustada aos dados, mas também em conformidade com os princípios físicos subjacentes. Com isso, torna-se possível resolver problemas diretos e inversos mesmo em cenários com dados escassos ou ruidosos, superando limitações típicas de modelos puramente baseados em dados (RAISSI; PERDIKARIS; KARNIADAKIS, 2019). Essa formulação evidencia como redes neurais podem ser treinadas para satisfazer EDPs por meio de técnicas de diferenciação automática (GOODFELLOW et al., 2016), consolidando uma nova classe de métodos numéricos baseados em aprendizado de máquina para modelagem científica.

Para compreender como as PINNs integram o conhecimento físico ao treinamento, é necessário explorar sua formulação matemática. Inicialmente, a construção dessas redes envolve a definição dos resíduos associados às EDPs e a composição da função de perda, que orienta o processo de otimização. Essa função de perda combina diferentes termos que asseguram a aderência da solução aos dados e às restrições físicas: a perda associada à condição inicial ($\mathcal{L}$ic), a perda relativa ao resíduo das EDPs ($\mathcal{L}$r), a perda relacionada às condições de fronteira ($\mathcal{L}$b) e a perda de dados ($\mathcal{L}$dt). Esta última é computada com base nos resultados numéricos obtidos, por exemplo, por meio do MVF ou nos dados experimentais relacionados ao fenômeno estudado.

De forma a elucidar o treinamento desse tipo de rede neural, podemos considerar, de forma geral, EDPs que seguem a estrutura:

$$\frac{\partial u}{\partial t} + \mathcal{N}[u] + g(t) = 0, \quad t \in [0, T],\ x \in \Omega, \tag{2.9}$$

com condições iniciais e de contorno dadas por:

$$u(0, x) = h(x), \quad x \in \Omega, \tag{2.10}$$

$$\mathcal{B}[u] = 0, \quad t \in [0, T], \ x \in \partial\Omega. \tag{2.11}$$

Para aproximar essa solução desconhecida, consideramos uma rede neural profunda parametrizada, denotada por $u_\theta(t, x)$, onde θ representa o conjunto de parâmetros ajustáveis da rede, tais como os pesos e os vieses.

Com essa representação, é possível definir os resíduos da EDP como (RAISSI; PERDIKARIS; KARNIADAKIS, 2019):

$$\zeta_\theta(t, x) = \frac{\partial u_\theta}{\partial t}(t, x) + \mathcal{N}[u_\theta](t, x) + g(t), \tag{2.12}$$

onde os pontos (t, x) pertencem ao domínio de interesse. O termo ζ_θ mede os resíduos da predição da rede neural em relação às equações diferenciais em pontos específicos do domínio, sendo, portanto, fundamental para o cálculo da função de perda informada pela física.

As derivadas parciais presentes na formulação dos resíduos, como $\frac{\partial u_\theta}{\partial t}$ e os termos diferenciais do operador $\mathcal{N}[u_\theta]$, são computadas por meio de diferenciação automática, uma técnica que permite a avaliação exata dos gradientes ao longo do grafo computacional da rede. Essa abordagem difere da diferenciação numérica tradicional, uma vez que não depende da aproximação de derivadas por meio de incrementos finitos em torno dos pontos de avaliação. No caso das PINNs, tal propriedade é essencial, pois permite a imposição direta das equações diferenciais no processo de treinamento, sem a necessidade de discretização do domínio, o que contribui para a obtenção de soluções consistentes com as leis físicas relacionadas ao fenômeno estudado.

Dessa forma, o modelo informado por física é treinado por meio da minimização da seguinte função de perda composta (RAISSI; PERDIKARIS; KARNIADAKIS, 2019):

$$\mathcal{L}(\theta) = \tau\, \mathcal{L}_{\mathrm{ic}}(\theta) + \kappa\, \mathcal{L}_{\mathrm{r}}(\theta) + \epsilon \mathcal{L}_{\mathrm{b}}(\theta) + \nu\, \mathcal{L}_{\mathrm{dt}}(\theta). \tag{2.13}$$

A perda associada à condição inicial, $\mathcal{L}_{\mathrm{ic}}$, quantifica o erro entre a predição da rede e os valores impostos pela condição inicial. A perda $\mathcal{L}_{\mathrm{r}}$ corresponde ao erro nos resíduos das equações diferenciais no interior do domínio. Já $\mathcal{L}_{\mathrm{b}}$ avalia a violação das condições de fronteira, enquanto $\mathcal{L}_{\mathrm{dt}}$ refere-se a perda de dados e mede o erro da rede em pontos para os quais há dados disponíveis, neste caso, provenientes do MVF. Já τ, κ, ϵ e ν são hiperparâmetros que ponderam a contribuição de cada termo na função de perda total.

Quando avaliadas por meio do Erro Quadrático Médio (*Root Mean Squared Error* - RMSE), as funções de perda podem ser expressas da seguinte forma:

$$\mathcal{L}_{\mathrm{ic}}(\theta) = \frac{1}{N_{\mathrm{ic}}} \sum_{i=1}^{N_{\mathrm{ic}}} \sqrt{(u_\theta(0, x_{ic}^i) - h(x_{ic}^i))^2}, \tag{2.14}$$

$$\mathcal{L}_{\mathrm{r}}(\theta) = \frac{1}{N_{\mathrm{r}}} \sum_{i=1}^{N_{\mathrm{r}}} \sqrt{(\zeta_\theta(t_r^i, x_r^i))^2}, \tag{2.15}$$

$$\mathcal{L}_{\mathrm{b}}(\theta) = \frac{1}{N_{\mathrm{b}}} \sum_{i=1}^{N_{\mathrm{b}}} \sqrt{(\mathcal{B}[u_\theta](t_b^i, x_b^i))^2}, \tag{2.16}$$

$$\mathcal{L}_{\mathrm{dt}}(\theta) = \frac{1}{N_{\mathrm{dt}}} \sum_{i=1}^{N_{\mathrm{dt}}} \sqrt{(u_\theta(t_{dt}^i, x_{dt}^i) - u(t_{dt}^i, x_{dt}^i))^2}, \tag{2.17}$$

onde os conjuntos de pontos $\{x_i^{\mathrm{ic}}\}_{i=1}^{N_{\mathrm{ic}}}$, $\{t_i^{\mathrm{b}}, x_i^{\mathrm{b}}\}_{i=1}^{N_{\mathrm{b}}}$, $\{t_i^{\mathrm{r}}, x_i^{\mathrm{r}}\}_{i=1}^{N_{\mathrm{r}}}$ e $\{t_i^{\mathrm{dt}}, x_i^{\mathrm{dt}}\}_{i=1}^{N_{\mathrm{dt}}}$ referem-se, respectivamente, aos pontos associados às condições iniciais, às condições de contorno, aos pontos de resíduo utilizados para impor as equações diferenciais e aos pontos da malha empregados para calcular a solução do MVF.

3 MÉTODOS

Este trabalho propõe uma abordagem comparativa para a simulação da dinâmica espaço-temporal de concentrações biológicas, avaliando o desempenho de três métodos distintos: o Método dos Volumes Finitos (MVF), redes neurais (NN) convencionais e redes neurais informadas por física (PINN). Para atingir tal objetivo, o método adotado consiste em três etapas principais:

- simulação do modelo de base por MVF;
- treinamento das redes neurais com diferentes densidades de dados;
- análise comparativa dos tempos de treinamento, inferência e precisão das soluções obtidas.

Inicialmente, o modelo foi discretizado e simulado pelo MVF em sua versão serial, implementada para execução em CPU. A discretização foi realizada no domínio espaço-temporal de interesse e o tempo de execução da simulação foi registrado. Em seguida, a mesma formulação numérica foi paralelizada para execução em GPU, explorando a decomposição espacial do domínio e a sincronização entre *threads* a cada passo de tempo. Com isso, foi possível comparar a aceleração obtida por paralelização em relação à implementação sequencial, tomando como base as mesmas condições iniciais e parâmetros do modelo.

A partir dos dados gerados pelo MVF serial, redes neurais foram treinadas para aproximar a solução do sistema dinâmico. Duas abordagens foram avaliadas: uma rede neural tradicional treinada apenas com dados (NN) e uma PINN. Para avaliar a sensibilidade dessas redes à disponibilidade de dados, o conjunto de amostras temporais utilizadas no treinamento foi sistematicamente reduzido, e os respectivos tempos de treinamento e inferência foram registrados. A inferência consistiu na previsão de toda a malha espaço-temporal, utilizando como entrada apenas os valores de tempo e espaço. Ambos os tempos de inferência e treinamento foram medidos também em GPU.

Admite-se que há um viés metodológico nesta comparação, uma vez que ambas as redes foram avaliadas utilizando a mesma topologia, sem a realização de uma busca de hiperparâmetros específica e independente para cada abordagem. Idealmente, redes PINNs e NNs deveriam ser comparadas em suas respectivas arquiteturas otimizadas, obtidas, por exemplo, por meio de uma busca em grade separada. No entanto, essa escolha metodológica foi intencional: o objetivo principal do experimento é isolar o impacto da imposição das restrições físicas no treinamento — característica fundamental das PINNs — em contraste com redes puramente orientadas a dados. Assim, optou-se por manter a

topologia fixa como forma de controlar outras variáveis e enfatizar a contribuição estrutural das restrições físicas.

Este capítulo apresenta o método em detalhes, iniciando pela apresentação do modelo matemático empregado para representar, de modo simplificado, a resposta do sistema imune contra um patógeno genérico.

3.1 MODELO DO SISTEMA IMUNE

O modelo matemático utilizado para descrever a fisiopatologia da formação de edema foi introduzido em um estudo anterior (REIS et al., 2019). Esse modelo incorpora o componente inflamatório, que detalha a interação entre um patógeno e o sistema imunológico humano. Além das dinâmicas inflamatórias, o modelo adota uma formulação poroelástica baseada na teoria de Biot, com o intuito de considerar a deformação mecânica do tecido simulado. Subsequentemente, tal modelo foi adaptado para investigar os efeitos da miocardite infecciosa sobre o coração (REIS et al., 2018). Ele integra a resposta imune à formação de edema extracelular. No entanto, este trabalho foca exclusivamente no componente da resposta do sistema imunológico. Consequentemente, o componente mecânico do modelo não fez parte das simulações apresentadas neste trabalho. O modelo representa o tecido como um meio poroso saturado por fluido e caracteriza o patógeno da seguinte forma:

$$\begin{cases} \dfrac{\partial\left(\phi C_p\right)}{\partial t} = \nabla \cdot \left(D_b \nabla C_p\right) - r_b + q_b \text{ em } \Omega \times I, \\ D_b \nabla C_p \cdot \mathbf{n} = 0 \ \partial\Omega \times I, \\ C_p(x,0) = C_{p0}(x) \ \Omega, \end{cases} \tag{3.1}$$

sendo $\Omega \subset \mathbb{R}^2$ e $I = [0, t_f) \subset \mathbb{R}^+$ o intervalo de tempo, $\mathbf{n}$ é o vetor normal orientado para fora da borda do domínio $\partial\Omega$, $C_p : \Omega \times I \to \mathbb{R}^+$ representa a concentração de patógenos no fluido intersticial, ϕ_f é a porosidade, D_b é o coeficiente de difusão dos patógenos através do fluido intersticial, q_b denota a reprodução dos patógenos, e r_b denota a morte dos patógenos devido à ação dos leucócitos. A difusão é definida como o espalhamento de partículas de regiões de maior concentração para regiões de menor concentração. Os termos q_b e r_b correspondem, respectivamente, às fontes de patógenos e à fagocitose de patógenos pelos leucócitos, sendo definidos por:

$$\begin{aligned} q_b &= C_p C_p, \\ r_b &= \lambda_{nb} C_l C_p, \end{aligned} \tag{3.2}$$

onde C_p é a taxa de crescimento dos patógenos no tecido intersticial, C_l é a concentração de leucócitos no fluido intersticial, e λ_{nb} representa a taxa de fagocitose exercida pelos leucócitos.

O modelo diferencial que representa a dinâmica dos leucócitos é dado por:

$$\begin{cases} \frac{d(\phi_f C_l)}{dt} = \nabla \cdot (D_n \nabla C_l - \chi_{nb} C l \nabla C_p) - r_n + q_n, \quad \text{em } \Omega \times I, \\ (D_n \nabla C_l - \chi_{nb} C_l \nabla C_p) \cdot \mathbf{n} = 0 \quad \text{em } \partial\Omega \times I, \\ C_l(x, 0) = C_{l0}(x), \end{cases} \tag{3.3}$$

sendo $C_l : \Omega \times I \to \mathbb{R}^+$ a concentração de leucócitos no fluido intersticial, D_n o coeficiente de difusão dos leucócitos através do tecido intersticial, χ_{nb} a taxa de quimiotaxia dos leucócitos, q_n a fonte de leucócitos, que representa o extravasamento de leucócitos da corrente sanguínea para o interstício, e r_n denota a morte dos leucócitos, tanto por apoptose quanto por fagocitose de patógenos. Os termos mencionados são definidos como:

$$q_n = \gamma_n C_p (C_{n,\max} - C_l), \tag{3.4}$$

$$r_n = \lambda_{bn} C_l C_p + \mu_n C_l, \tag{3.5}$$

onde $C_{n,\max}$ é a concentração máxima de leucócitos na corrente sanguínea, e γ_n representa a permeabilidade dos leucócitos à parede microvascular capilar. Nesse contexto, λ_{bn} denota a taxa de morte celular induzida após a fagocitose, enquanto μ_n refere-se à taxa de decaimento natural dos leucócitos, uma vez que parte dessas células é programada para morrer logo após sair da circulação sanguínea.

As condições de contorno adotadas neste estudo são do tipo Neumann homogêneo, isto é, impõem fluxo na fronteira do domínio. Essa escolha reflete a hipótese de isolamento funcional da região simulada em relação ao restante do tecido miocárdico, assumindo que não há troca significativa de patógenos ou leucócitos com áreas adjacentes durante o intervalo de tempo considerado. Tal suposição é particularmente pertinente no contexto de lesões inflamatórias focais típicas da miocardite. Ao restringir o modelo a um domínio fechado, preserva-se a coerência física da simulação, ao mesmo tempo em que se reduzem os graus de liberdade do sistema, permitindo uma análise mais controlada e precisa da dinâmica espaço-temporal interna.

Os valores dos parâmetros utilizados em todas as simulações são os mesmos apresentados na Tabela 1, que foram definidos com base na simulação do tecido miocárdico.

3.2 SUPOSIÇÕES DO MODELO

Embora o processo inflamatório e sua resolução envolvam uma ampla gama de proteínas e células imunes, este trabalho foca especificamente em determinadas células do sistema imune inato, conhecidas como leucócitos, ou seja, os glóbulos brancos que atuam na defesa contra patógenos. Dentre essas células estão os macrófagos, células dendríticas, neutrófilos, eosinófilos, células T, células B e células natural killer, todas

Tabela 1 – Valores dos parâmetros utilizados nas Equações 3.1 e 3.3, com base em (REIS et al., 2019).

Nome	Valor
Porosidade (ϕ_f)	0.2
Coeficiente de difusão de patógenos (D_b)	0.005 $\frac{cm^2}{h}$
Taxa de reprodução de patógenos (c_p)	0.15 $\frac{1}{h}$
Taxa de fagocitose (λ_{nb})	1.8 $\frac{cm^2}{h \cdot 10^7 \ cell}$
Coeficiente de difusão de leucócitos (D_n)	5×10^{-05} $\frac{cm^2}{h}$
Taxa de quimiotaxia (χ_{nb})	0.1 $\frac{cm^5}{h \cdot 10^7 cell}$
Taxa de apoptose induzida (λ_{bn})	1.8 $\frac{cm^3}{h \cdot 10^{10} cell}$
Permeabilidade capilar para os leucócitos (γ_n)	0.1 $\frac{cm^3}{h \cdot 10^7 cell}$
Concentração máxima de leucócitos no sangue ($C_{n,max}$)	550.0 $cell$
Taxa de apoptose (μ_n)	0.2 $\frac{1}{h}$

capazes de combater diferentes tipos de patógenos, incluindo fungos, vírus, bactérias e parasitas de maior porte. O modelo proposto analisa os patógenos e os leucócitos em um meio poroso com propriedades homogêneas e isotrópicas, saturado por fluido intersticial, representando o tecido vivo. Assume-se que os leucócitos são capazes de eliminar os patógenos do interstício por mecanismos como a fagocitose, enquanto os patógenos, por sua vez, possuem capacidade de reprodução ou replicação. Uma vez que os leucócitos eliminam os patógenos do tecido, eles atuam na resolução da inflamação por meio da remoção das células recrutadas e na restauração da homeostase tecidual.

Este estudo modela o tecido como sendo o miocárdio, embora a estrutura conceitual adotada seja suficientemente geral para ser aplicada a outros tipos de tecidos. O modelo não contempla o comportamento ativo do músculo cardíaco, tampouco inclui os processos eletrofisiológicos ou as interações mecânicas entre o coração e os tecidos ou órgãos adjacentes. O principal desafio para a incorporação desses aspectos ao modelo poroelástico reside nas diferenças de escalas temporais envolvidas (LOURENÇO et al., 2022). A eletrofisiologia cardíaca e as contrações do coração ocorrem em escalas de tempo muito mais rápidas quando comparadas ao processo mais lento de formação do edema, que constitui o foco principal deste estudo.

3.3 DISCRETIZAÇÃO PELO MÉTODO DOS VOLUMES FINITOS

Para a obtenção de soluções aproximadas do sistema de equações diferenciais parciais que modela a interação entre patógenos e leucócitos, utilizou-se o Método dos Volumes Finitos (MVF). Esta abordagem consiste em integrar as equações de conservação sobre volumes de controle no domínio espacial, assegurando a conservação local das quantidades físicas.

Considerando um volume de controle centrado no ponto x_i, com faces localizadas em $x_{i-\frac{1}{2}}$ e $x_{i+\frac{1}{2}}$, a equação de conservação para a concentração de patógenos, C_p, é integrada

ao longo do volume:

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial(\phi_f C_p)}{\partial t}\, dx = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial}{\partial x}\left(D_b \frac{\partial C_p}{\partial x}\right) dx + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (q_b - r_b)\, dx. \tag{3.6}$$

A equação acima representa o balanço da concentração de patógenos dentro do volume de controle, sendo ϕ_f a porosidade, D_b o coeficiente de difusão, q_b a taxa de produção de patógenos, e r_b a taxa de remoção devido à ação dos leucócitos.

Aplicando o Teorema Fundamental do Cálculo ao termo difusivo e rearranjando os termos, obtém-se:

$$\phi_f \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial C_p}{\partial t}\, dx = D_b \frac{\partial C_p}{\partial x}\bigg|_{i+\frac{1}{2}} - D_b \frac{\partial C_p}{\partial x}\bigg|_{i-\frac{1}{2}} + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (q_b - r_b)\, dx. \tag{3.7}$$

Para a aproximação das integrais, utilizamos a regra do ponto médio, assumindo que as variáveis são aproximadamente constantes no interior do volume de controle. O tamanho do volume é dado por $h = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, resultando em:

$$h\phi_f \frac{\partial C_p}{\partial t} = D_b \frac{\partial C_p}{\partial x}\bigg|_{i+\frac{1}{2}} - D_b \frac{\partial C_p}{\partial x}\bigg|_{i-\frac{1}{2}} + h(q_b - r_b). \tag{3.8}$$

As derivadas espaciais nas interfaces são aproximadas por diferenças finitas centradas:

$$\frac{\partial C_p}{\partial x}\bigg|_{i+\frac{1}{2}} \approx \frac{C_{p_{(i+1)}} - C_{p_{(i)}}}{h}, \tag{3.9}$$

$$\frac{\partial C_p}{\partial x}\bigg|_{i-\frac{1}{2}} \approx \frac{C_{p_{(i)}} - C_{p_{(i-1)}}}{h}. \tag{3.10}$$

A discretização temporal é realizada por diferenças progressivas de primeira ordem, com passo de tempo k:

$$\frac{\partial C_p}{\partial t} \approx \frac{C^{n+1}_{p_{(i)}} - C^{n}_{p_{(i)}}}{k}. \tag{3.11}$$

Substituindo as aproximações acima na equação integral, obtemos:

$$h\phi_f \frac{C^{n+1}_{p_{(i)}} - C^{n}_{p_{(i)}}}{k} = D_b \left(\frac{C^{n}_{p_{(i+1)}} - C^{n}_{p_{(i)}}}{h} - \frac{C^{n}_{p_{(i)}} - C^{n}_{p_{(i-1)}}}{h} \right) + h(q_b - r_b). \tag{3.12}$$

Agrupando os termos e isolando $C^{n+1}_{p_{(i)}}$, temos a fórmula final explícita de atualização temporal:

$$C_{p_{(i)}}^{n+1} = \frac{kD_b}{h^2\phi_f}\left[C_{p_{(i+1)}}^{n} - 2C_{p_{(i)}}^{n} + C_{p_{(i-1)}}^{n}\right] + \frac{k}{\phi_f}(q_b - r_b) + C_{p_{(i)}}^{n}. \tag{3.13}$$

Essa formulação explícita do método dos volumes finitos permite a evolução temporal da concentração de patógenos a partir das condições iniciais e de contorno, incorporando os efeitos de difusão, produção e remoção pela resposta imune.

De modo análogo à equação dos patógenos, a equação que governa a concentração de leucócitos C_l também foi discretizada utilizando MVF. A equação de conservação é formulada sobre um volume de controle centrado no ponto x_i, com faces nos pontos $x_{i-\frac{1}{2}}$ e $x_{i+\frac{1}{2}}$. A equação contínua é dada por:

$$\int_{x_{i-\frac{1}{2}}}^{x_i+\frac{1}{2}} \frac{\partial(\phi_f C_l)}{\partial t}\,dx = \int_{x_{i-\frac{1}{2}}}^{x_i+\frac{1}{2}} \frac{\partial}{\partial x}\left(D_n\frac{\partial C_l}{\partial x} - \chi_{nb}C_l\frac{\partial C_p}{\partial x}\right)dx + \int_{x_{i-\frac{1}{2}}}^{x_i+\frac{1}{2}}(q_n - r_n)\,dx. \tag{3.14}$$

Neste modelo, D_n é o coeficiente de difusão dos leucócitos, χ_{nb} é o coeficiente de quimiotaxia, q_n representa a fonte de leucócitos (extravasamento), e r_n corresponde à sua remoção (apoptose ou fagocitose de patógenos).

A aplicação do Teorema Fundamental do Cálculo aos termos difusivos e advectivos permite expressar os fluxos nas interfaces como:

$$\begin{aligned}\int_{x_{i-\frac{1}{2}}}^{x_i+\frac{1}{2}} \phi_f\frac{\partial C_l}{\partial t}\,dx = \left(D_n\frac{\partial C_l}{\partial x} - \chi_{nb}C_l\frac{\partial C_p}{\partial x}\right)\Bigg|_{i+\frac{1}{2}} - \left(D_n\frac{\partial C_l}{\partial x} - \chi_{nb}C_l\frac{\partial C_p}{\partial x}\right)\Bigg|_{i-\frac{1}{2}} + \\ \int_{x_{i-\frac{1}{2}}}^{x_i+\frac{1}{2}}(q_n - r_n)\,dx.\end{aligned} \tag{3.15}$$

Aplicando a regra do ponto médio para as integrais, obtemos:

$$h\phi_f\frac{\partial C_l}{\partial t} = \left(D_n\frac{\partial C_l}{\partial x} - \chi_{nb}C_l\frac{\partial C_p}{\partial x}\right)\Bigg|_{i+\frac{1}{2}} - \left(D_n\frac{\partial C_l}{\partial x} - \chi_{nb}C_l\frac{\partial C_p}{\partial x}\right)\Bigg|_{i-\frac{1}{2}} + h(q_n - r_n). \tag{3.16}$$

As derivadas espaciais são aproximadas por diferenças progressivas de forma análoga às equações (3.9), (3.10) e (3.11). Neste caso, temos que a equação discretizada resultante torna-se:

$$\begin{aligned}h\phi_f\frac{C_{l_{(i)}}^{n+1} - C_{l_{(i)}}^{n}}{k} = &\frac{D_n}{h}\left(C_{l_{(i+1)}}^{n} - 2C_{l_{(i)}}^{n} + C_{l_{(i-1)}}^{n}\right) \\ &- \frac{\chi_{nb}}{h}\left[C_{l_\delta}^{n}(C_{p_{(i+1)}}^{n} - C_{p_{(i)}}^{n}) - C_{l_\varsigma}^{n}(C_{p_{(i)}}^{n} - C_{p_{(i-1)}}^{n})\right] \\ &+ h(q_n - r_n).\end{aligned} \tag{3.17}$$

Isolando $C_{l_{(i)}}^{n+1}$, obtemos:

$$\begin{aligned} C_{l_{(i)}}^{n+1} = C_{l_{(i)}}^{n} &+ \frac{kD_n}{h^2\phi_f}\left(C_{l_{(i+1)}}^{n} - 2C_{l_{(i)}}^{n} + C_{l_{(i-1)}}^{n}\right) \\ &- \frac{k\chi_{nb}}{h^2\phi_f}\left[C_{l_\delta}^{n}(C_{p_{(i+1)}}^{n} - C_{p_{(i)}}^{n}) - C_{l_\zeta}^{n}(C_{p_{(i)}}^{n} - C_{p_{(i-1)}}^{n})\right] \\ &+ \frac{k}{\phi_f}(q_n - r_n). \end{aligned} \tag{3.18}$$

A equação acima apresenta termos de difusão e advecção quimiotática, sendo este último responsável pelo transporte direcionado de leucócitos em resposta ao gradiente de concentração de patógenos. Para tratar adequadamente esse termo convectivo, aplicou-se o esquema de estabilização a montante, no qual os índices δ e ζ são escolhidos com base no sinal do gradiente de concentração de C_p:

$$\delta = \begin{cases} i, & \text{se } \left.\frac{\partial C_p}{\partial x}\right|_{i+\frac{1}{2}} \geq 0, \\ i+1, & \text{se } \left.\frac{\partial C_p}{\partial x}\right|_{i+\frac{1}{2}} < 0, \end{cases} \tag{3.19}$$

$$\zeta = \begin{cases} i, & \text{se } \left.\frac{\partial C_p}{\partial x}\right|_{i-\frac{1}{2}} < 0, \\ i-1, & \text{se } \left.\frac{\partial C_p}{\partial x}\right|_{i-\frac{1}{2}} \geq 0. \end{cases} \tag{3.20}$$

A formulação considera condições de contorno do tipo Neumann homogêneo (fluxo nulo), e as condições iniciais assumem ausência de leucócitos no tecido, ou seja, $C_l(x, 0) = 0$ e uma concentração inicial δ_b de patógenos em um ponto ou região específica.

3.3.1 Critério de estabilidade numérica

Para garantir a estabilidade do esquema numérico explícito adotado, especialmente em problemas com múltiplos mecanismos de transporte, como difusão e quimiotaxia, foi aplicada a *condição de Courant–Friedrichs–Lewy* (CFL)(LEVEQUE, 2007). Esta condição impõe uma restrição conjunta ao passo de tempo k e ao espaçamento espacial h, de modo a assegurar que a propagação numérica da informação não ultrapasse o domínio de dependência física do problema.

No presente modelo, que inclui termos difusivos, advectivos e reativos, a condição de CFL foi formulada considerando as contribuições de cada termo da equação. Em sua forma mais completa para o caso unidimensional, adota-se a seguinte expressão:

$$\frac{kc_x}{h} + \frac{2kD}{h^2} + \frac{kr}{2} < 1, \tag{3.21}$$

em que c_x representa a velocidade característica associada à quimiotaxia, D é o coeficiente de difusão e r corresponde a um parâmetro de reação, assumido aqui como constante por linearização. O primeiro termo limita o avanço das partículas pelo transporte direcionado;

o segundo assegura a estabilidade da difusão, enquanto o último leva em conta o efeito estabilizante (ou desestabilizante) da reação. A violação dessa condição pode conduzir a oscilações e à instabilidade da simulação. Assim, o passo de tempo k foi escolhido de modo a satisfazer rigorosamente essa desigualdade, garantindo a convergência das soluções obtidas.

3.4 REDES NEURAIS

A Figura 4 ilustra a arquitetura da rede neural utilizada neste estudo. A camada de entrada é composta por dois neurônios, correspondentes ao instante de tempo t e o ponto no espaço x para os quais queremos prever a solução latente da equação. A camada de saída também contém dois neurônios, associados às concentrações de patógenos (C_p) e de leucócitos (C_l).

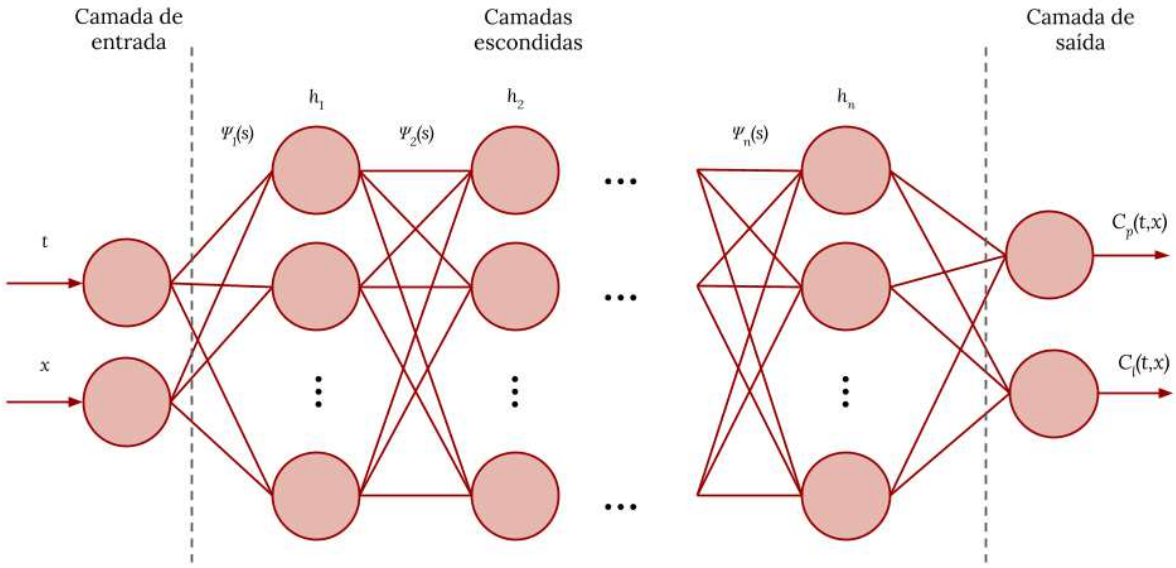


Figura 4 – Arquitetura da rede neural profunda empregada para o treinamento do modelo PINN.

Neste trabalho, adotou-se a função tangente hiperbólica como função de ativação de todos os neurônios, visto que esta é a função mais utilizada na literatura. Tomando como base a j-ésima camada oculta, podemos definir essa função de ativação por:

$$\psi_j(l) = \frac{\sinh(l)}{\cosh(l)}, \tag{3.22}$$

onde l representa a combinação linear dos valores de saída dos neurônios da camada anterior, ponderadas por seus respectivos pesos e acrescida de um termo de viés. Essa combinação é dada por:

$$l = \left(\sum_{i=1}^{N_{j-1}} w_{(j,i)} s_{(j,i)} \right) + b_j, \tag{3.23}$$

em que $s_{(j,i)}$ denota o valor de ativação do i-ésimo neurônio da camada anterior, $w_{(j,i)}$ é o peso da conexão correspondente e b_j é o viés adicionado ao neurônio da camada atual. A escolha da tangente hiperbólica justifica-se por sua natureza suave e centrada em zero, o que pode facilitar o processo de otimização durante o treinamento da rede, além de favorecer a representação da solução latente.

Com o intuito de encontrar a arquitetura que melhor resolve o compromisso entre custo de treino e acurácia, foi conduzido um processo de busca em grade sobre os hiperparâmetros estruturais da rede, número de camadas ocultas e quantidade de neurônios por camada, bem como sobre os os coeficientes de momento β_1 e β_2 do otimizador ADAM apresentados na Equação 2.4. O processo de otimização está intrinsecamente vinculado à busca, uma vez que cada ponto da grade corresponde a um modelo treinado até a convergência, sendo então avaliado quanto à sua acurácia, com base no RMSE e à sua eficiência computacional, considerando a aceleração em relação ao MVF serial. Assim, a variação dos parâmetros de treinamento justifica-se pela sua influência direta sobre o erro. A escolha final da arquitetura e dos parâmetros do otimizador foi baseada no melhor desempenho conjunto nesses dois critérios. Uma vez identificada a configuração ótima para as PINNs, a mesma foi empregada para o treinamento das NNs, o que possibilita uma análise isolada do impacto da imposição explícita das restrições físicas sobre o processo de aprendizado.

3.5 IMPLEMENTAÇÃO

A formulação diferencial das equações governantes foi discretizada utilizando o MVF com esquema explícito, adotando-se um passo espacial $h = 0{,}01$ e um passo temporal $k = 0{,}001$. O domínio considerado corresponde a um tecido de 1,0 cm de comprimento, simulado ao longo de um intervalo temporal de 1,5 horas. Assim, foram utilizados 10^2 pontos no espaço e $1{,}5 \times 10^3$ passos no tempo, totalizando $1{,}5 \times 10^5$ amostras espaço-temporais que compõem a base de dados de referência utilizada para o treinamento e validação das redes neurais. A malha adotada é regular, com espaçamento uniforme tanto no espaço quanto no tempo.

Adicionalmente, o termo fonte responsável pela entrada de leucócitos no tecido foi ponderado por uma matriz booleana que representa a distribuição espacial dos vasos sanguíneos. Esses vasos foram modelados como pontos de entrada fixos ao longo do domínio, definidos de forma aleatória no início da simulação e correspondentes a 10% dos pontos da malha espacial. Essa abordagem permitiu representar de forma simplificada

parte da heterogeneidade anatômica do tecido e conferir maior realismo à distribuição das fontes inflamatórias.

Levando em consideração a natureza sequencial do esquema explícito adotado, a versão do MVF implementada para execução em GPU computa a evolução temporal de forma iterativa, em que cada passo no tempo depende diretamente do estado anterior da malha. Por outro lado, a atualização dos pontos no espaço, dentro de um mesmo instante de tempo, é independente e pôde ser paralelizada. Essa paralelização foi viabilizada por meio da biblioteca `Numba` (LAM et al., 2024), que oferece suporte à geração de código otimizado para GPU através da interface com CUDA.

Mais especificamente, foram utilizadas as diretivas `@cuda.jit` do `Numba` para compilar funções em linguagem intermediária capaz de ser executada diretamente nos núcleos CUDA da GPU. O domínio espacial foi mapeado para uma grade unidimensional de *threads*, de forma que cada *thread* ficou responsável pelo processamento de um ponto da malha espacial. Ao final de cada passo temporal, as matrizes que armazenam os estados das variáveis foram sincronizadas, respeitando a dependência temporal do esquema de resolução, ou seja, os resultados de cada ponto em t^{n+1} só são calculados após a conclusão das atualizações em t^n. Esse controle de fluxo é fundamental para garantir a consistência da solução numérica no tempo, dado que o método explícito não permite avaliação simultânea entre diferentes instantes temporais.

Cabe ressaltar que, ao utilizar `Numba`, a compilação JIT (*Just-In-Time*) dos *kernels* CUDA ocorre na primeira execução de cada função decorada com `@cuda.jit`, o que introduz um *overhead* inicial de tempo. No entanto, como essa etapa ocorre apenas uma vez e não afeta o tempo de simulação subsequente, optou-se por desconsiderar esse custo inicial no cálculo da aceleração computacional reportada neste trabalho, a fim de refletir com mais fidelidade o desempenho real da simulação após a compilação das rotinas e aproximar o resultado.

Durante o processo de busca em grade, alguns hiperparâmetros foram mantidos fixos. A taxa de aprendizado foi definida como $\alpha = 0{,}001$, enquanto o termo de estabilização numérica foi fixado em $\varepsilon = 10^{-8}$, conforme empregado na equação de atualização dos parâmetros da rede (Equação (2.8)). Para o treinamento das PINNs, utilizaram-se os coeficientes $\tau = 1$, $\kappa = 5$, $\epsilon = 1$ e $\nu = 1$ na composição da função de perda total (Equação (2.13)), ponderando os termos associados às condições iniciais, aos resíduos das equações diferenciais, às condições de contorno e aos dados de referência, respectivamente.

O conjunto de dados de referência, gerado por meio do MVF, foi dividido em duas partes: 90% das amostras foram destinadas ao treinamento e 10% à validação. As *batchs* utilizadas durante o treino foram construídas exclusivamente com os 90% referentes ao conjunto de treinamento. Em cada época de treinamento, a rede foi apresentada a dez *batchs* compostas por 10^4 amostras selecionadas aleatoriamente a partir dos dados de

referência. Cada ponto foi utilizado uma única vez por época, o que garante que todos os dados disponíveis foram processados a cada ciclo completo de treinamento. Portanto, os pontos empregados na perda de dados correspondem aos vértices da malha regular gerada no domínio espaço-temporal.

Os demais termos da função de perda (condição inicial, condição de contorno e resíduo da equação diferencial) também foram avaliados com 10^4 amostras por *batch*. Esses pontos foram sorteados de forma independente, por amostragem uniforme no domínio específico de cada termo. Especificamente, os pontos referentes à condição inicial foram extraídos do instante $t = 0$; os pontos de contorno da borda $\partial\Omega$ ao longo do tempo; e os pontos do resíduo da equação foram amostrados no interior do domínio espaço-temporal $\Omega \times [0, T]$.

O número total de épocas de treinamento foi fixado em 10^4, sendo cada época composta por uma iteração completa do algoritmo de otimização, na qual os parâmetros da rede foram atualizados com base nos gradientes computados a partir das dez *batchs*.

Ademais, ressalta-se que os resultados apresentados nos gráficos e mapas de calor foram obtidos a partir dos mesmos pontos utilizados no cálculo da perda de dados. Dessa forma, as métricas de erro reportadas não refletem a capacidade de generalização das redes para novas condições, mas sim sua aptidão em replicar as soluções latentes geradas pelo MVF dentro do mesmo domínio de treinamento.

As redes neurais foram implementadas com base no *framework* `Pinn-Torch`, desenvolvido de forma aberta (WERNECK, 2025). Esse *framework* é construído sobre a biblioteca `PyTorch`2.2.1(PASZKE, 2024), a qual oferece suporte nativo à diferenciação automática, recurso utilizado para o cálculo eficiente dos gradientes em relação às variáveis de entrada e aos parâmetros treináveis θ. Todo o código-fonte foi desenvolvido em `Python` 3.9.18.

Com o intuito de garantir transparência e reprodutibilidade, todos os materiais produzidos, incluindo os *scripts*, rotinas de pré- e pós-processamento, bem como os *notebooks* utilizados na condução dos experimentos, encontram-se disponíveis em repositório público no GitHub (FERNANDES, 2025).

4 RESULTADOS

Este capítulo detalha os resultados obtidos com a solução do modelo descrito na Seção 3.1, utilizando três abordagens distintas: MVF (nas versões serial e com paralelização via CUDA), redes neurais convencionais (NN) e redes neurais informadas por física (PINN). O principal objetivo foi investigar a arquitetura das PINNs por meio de busca em grade (*grid search*) e comparar o desempenho de inferência entre as abordagens, avaliando tanto a acurácia quanto a eficiência computacional. Adicionalmente, especificamente para o treinamento das redes, analisamos o efeito da inclusão de restrições físicas na acurácia e na eficiência computacional.

No que diz respeito à investigação da arquitetura PINN, analisamos o impacto de diferentes configurações (como o número de camadas ocultas, a quantidade de neurônios por camada e as funções de ativação) na precisão e no tempo de execução dos modelos. O objetivo era identificar a arquitetura que oferecesse o melhor equilíbrio entre desempenho e custo computacional.

Cada experimento foi repetido 33 vezes, e as métricas de erro e os tempos de execução (tanto de treinamento quanto de inferência) foram registrados. Os resultados são apresentados em termos de média e desvio padrão.

As acelerações computacionais reportadas para os experimentos de inferência foram calculadas tendo como referência o tempo de execução da abordagem MVF em sua versão sequencial em CPU.

4.1 AMBIENTE COMPUTACIONAL

Os experimentos foram executados, inicialmente, em modo sequencial (*single-thread*), em um ambiente computacional composto por um processador *AMD EPYC 7713* com 128 núcleos físicos. Cada núcleo dispõe de 64 KB de *cache* L1 para dados, 64 KB de *cache* L1 para instruções, 512 KB de *cache* L2 unificado e compartilha 32 MB de *cache* L3 com outros 7 núcleos. A máquina possui, ainda, 528 GB de memória RAM.

Posteriormente, os mesmos experimentos foram executados em GPU, utilizando uma placa *NVIDIA A100* com 80 GB de memória HBM2e, com o objetivo de explorar a paralelização massiva oferecida pelos *CUDA cores*. Todas as redes neurais, tanto as redes tradicionais quanto as PINNs, foram treinadas e avaliadas exclusivamente nessa GPU, de modo a mensurar os ganhos de desempenho em relação à execução em CPU.

4.2 BUSCA EM GRADE DA ARQUITETURA PINN

A busca em grade é uma técnica fundamental para a otimização da arquitetura de modelos, especialmente no contexto de aprendizado profundo. Ela consiste em explorar

sistematicamente um conjunto pré-definido de hiperparâmetros. No trabalho em questão, o número de camadas ocultas e a quantidade de neurônios por camada foram o alvo deste estudo para identificar a combinação que proporciona o melhor desempenho para o problema proposto.

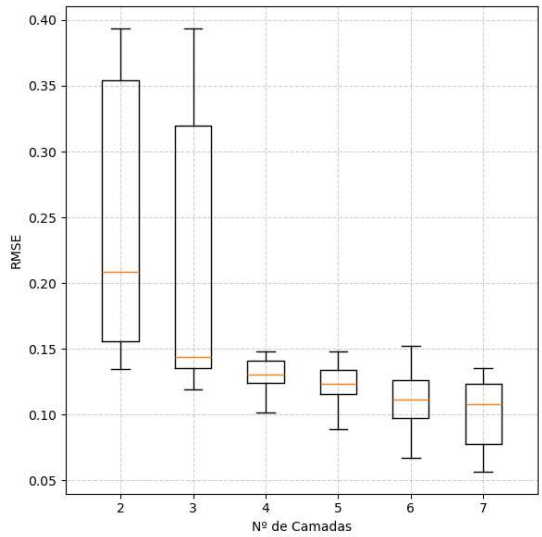
Ao avaliar todas as possíveis configurações, essa abordagem permite encontrar a arquitetura que melhor equilibra precisão e eficiência computacional. Este método torna-se essencial, pois o desempenho das redes neurais é altamente sensível aos hiperparâmetros, sendo que uma arquitetura otimizada pode resultar em melhorias significativas nos resultados do modelo.

Neste estudo, investigaram-se arquiteturas compostas por 2 a 7 camadas ocultas, com a quantidade de neurônios por camada variando entre 16 e 64. Além disso, estudamos também os seguinte parâmetros do ADAM, β_1 variando de 0,6 a 0,9 e β_1 variando de 0,99 a 0,9999. Esses intervalos foram escolhidos com base nos valores mais utilizados na literatura e em testes empíricos aplicados ao problema proposto. A função de ativação utilizada foi a `Tanh` (tangente hiperbólica) apresentada na Equação 3.22.

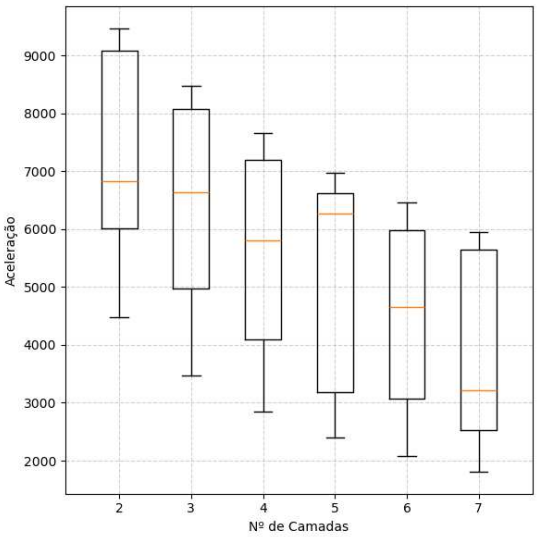
A combinação dessas configurações gerou um total de 600 arquiteturas distintas das quais 2 apresentaram problemas com o desaparecimento do gradiente e foram excluídas da análise. O desempenhos das arquiteturas válidas estão ilustrados nas Figuras 5 e 6. Os critérios utilizados para avaliar os modelos foram o RMSE, que representa o erro quadrático médio, e a aceleração média, que representa o ganho de desempenho computacional com a adoção do método proposto.

Na Figura 5a, observa-se a distribuição dos valores de RMSE em função do número de camadas ocultas utilizadas na arquitetura da rede neural. As arquiteturas com 5, 6 e 7 camadas apresentam medianas de erro bastante próximas, sugerindo que, a partir de cinco camadas, o modelo atinge um patamar de desempenho relativamente estável em termos de acurácia. No entanto, nota-se que a distribuição associada à arquitetura com 7 camadas apresenta uma dispersão ligeiramente maior, evidenciada pela amplitude interquartil mais extensa e pela presença de valores extremos mais afastados. Em contraponto, arquiteturas com número inferior de camadas, como 3 e 4, apresentam mediana de RMSE mais elevada, denotando menor capacidade de representação da dinâmica do sistema modelado.

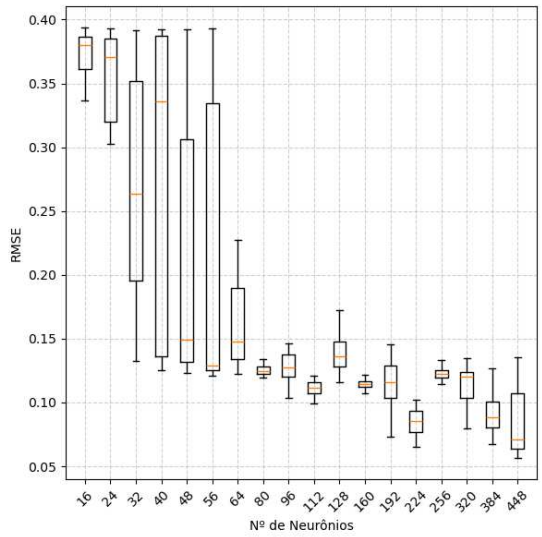
Ademais, na Figura 5b, observa-se que as redes com 5 camadas ocultas apresentam, em média, os maiores valores de aceleração, calculada com base na execução sequencial do MVF, sugerindo um potencial significativo de ganho em desempenho. No entanto, essa arquitetura exibe também a maior dispersão relativa entre os quartis, evidenciando uma variabilidade elevada nos resultados e indicando um desempenho menos consistente em termos de tempo de inferência. Por outro lado, os modelos com 6 e 7 camadas, apesar de apresentarem medianas de aceleração inferiores, possuem distribuições que se sobrepõem àquela de 5 camadas, especialmente dentro dos intervalos interquartis. Isso sugere que,



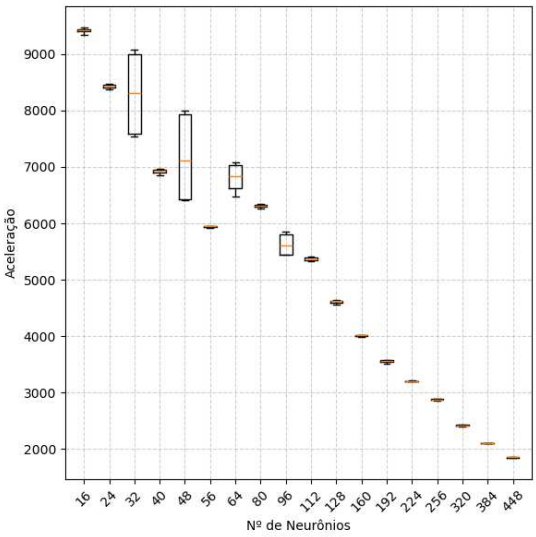
(a) RMSE em função do número de camadas.



(b) Aceleração em função do número de camadas.



(c) RMSE em função do número de neurônios.



(d) Aceleração em função do número de neurônios.

Figura 5 – Análise estatística da influência do número de camadas e neurônios sobre o RMSE e a aceleração.

dependendo da quantidade de neurônios por camada, essas arquiteturas podem atingir níveis de aceleração comparáveis. Já arquiteturas com menor número de camadas (2 a 4) tendem a apresentar medianas elevadas, mas com menor capacidade de representar a dinâmica das soluções latentes. Uma vez que nosso objetivo é acelerar as simulações da resposta imune no coração e que as acelerações mínimas já representam um avanço significativo, nesta análise o erro torna-se o fator principal. Assim, o modelo com menor erro, sem comprometer desnecessariamente o tempo de execução, foi selecionado como a melhor opção para resolver o problema.

A Figura 5c apresenta a distribuição dos valores de RMSE em função do número total de neurônios na arquitetura da rede, considerando a soma dos neurônios em todas as camadas ocultas. Observa-se uma tendência clara de redução do erro à medida que o número total de neurônios aumenta, especialmente nas faixas inferiores (até cerca de 96 neurônios totais), onde ocorre uma queda significativa tanto da mediana quanto da variabilidade do RMSE.

A partir de aproximadamente 80 neurônios totais, os valores medianos tornam-se mais estáveis e baixos, sugerindo que a rede passa a ter capacidade representacional suficiente para descrever o fenômeno modelado. No entanto, nota-se que alguns valores de neurônios, como por exemplo em torno de 48, 64 ou mesmo 224, apresentam distribuições com variância mais elevada. Tal comportamento pode ser atribuído ao fato de que um mesmo número total de neurônios pode ser alcançado por diferentes configurações de topologia — variando o número de camadas e a distribuição de neurônios entre elas — o que afeta diretamente o desempenho da rede. Por exemplo, uma rede com duas camadas de 32 neurônios (totalizando 64) pode ter comportamento distinto de outra com quatro camadas de 16 neurônios. Assim, a dispersão observada em determinados pontos reflete não apenas a influência da capacidade total, mas também a diversidade estrutural associada a esse mesmo total. Além disso, podemos perceber que as distribuições com menores medianas e valores mínimos de RMSE são as PINNs com 224, 384 e 448 neurônios.

A Figura 5d exibe a variação da aceleração em função do número total de neurônios nas arquiteturas avaliadas. Observa-se uma tendência decrescente clara: redes com menor número de neurônios tendem a apresentar maiores valores de aceleração, evidenciando um tempo de execução mais eficiente. Em particular, arquiteturas com até 64 neurônios totais demonstram acelerações médias próximas a 7.000, com destaque para as configurações mais compactas (16 a 40 neurônios), que alcançam valores próximos ou superiores a 9.000.

À medida que o número total de neurônios nas redes neurais aumenta, observamos uma redução progressiva na aceleração, atingindo valores em torno de 2.000 para redes com 448 neurônios. Esse comportamento é um reflexo direto do aumento do custo computacional associado à crescente complexidade da rede. É importante notar que diferentes combinações de profundidade e largura da rede podem resultar no mesmo número total de neurônios. No entanto, essas arquiteturas distintas podem ter impactos variados na eficiência computacional, o que também se reflete no aumento da dispersão dos resultados nesta análise.

A partir dessa análise conjunta da Figura 5, verifica-se que a arquitetura composta por 7 camadas e 32 neurônios em cada camada — totalizando 224 neurônios — representa a melhor solução de compromisso entre precisão e desempenho computacional. Essa configuração obteve valores de RMSE comparáveis aos das arquiteturas com maior número total de neurônios, mantendo, ao mesmo tempo, uma aceleração computacional equiparável

a arquiteturas menos profundas.

No que tange à influência estrutural da topologia da rede, é importante destacar que o aumento do número de camadas (profundidade) e o aumento do número de neurônios (largura) impactam o modelo de formas distintas. A ampliação do número de camadas tende a favorecer a extração de características e a modelagem de relações mais complexas entre as variáveis (LECUN; BENGIO; HINTON, 2015), o que pode ser particularmente vantajoso em problemas com dinâmicas temporais ou espaciais complexas. No entanto, redes muito profundas são mais difíceis de treinar, podendo sofrer com problemas de instabilidade numérica, como o desaparecimento do gradiente, além de exigirem maior tempo de convergência (BENGIO; SIMARD; FRASCONI, 1994). Esse comportamento também pode ser observado neste trabalho, visto que o aumento do número de camadas reduziu o erro mediano das topologias. No entanto, observa-se que o intervalo interquartil do erro aumenta a partir de quatro camadas, indicando que essas topologias são mais complexas de treinar.

Por outro lado, o aumento do número de neurônios por camada incrementa a capacidade de representação da rede de maneira mais direta, ampliando o espaço funcional acessível ao modelo (CYBENKO, 1989). Contudo, esse aumento impacta fortemente o custo computacional por iteração, uma vez que o número de parâmetros cresce quadraticamente com a largura, especialmente em camadas densamente conectadas. Além disso, redes muito largas podem estar mais sujeitas ao sobreajuste, sobretudo em contextos com conjuntos de dados limitados (ZHANG et al., 2017).

Assim, os resultados indicam que, para o problema em questão, o aumento da profundidade (com uma quantidade moderada de neurônios por camada) foi mais eficaz do que simplesmente ampliar o número total de neurônios. Tal configuração proporcionou um modelo expressivo o suficiente para capturar a complexidade do sistema, mas sem incorrer nos elevados custos computacionais associados a redes excessivamente largas.

A Figura 6 ilustra esses resultados sob uma outra ótica, também levando em consideração a combinação de parâmetros do ADAM, além das diferentes quantidades de camadas ocultas e neurônios por camada. Isso pode ser percebido pelos múltiplos pontos com o mesmo número de camadas ou neurônios e mesma aceleração, mas com valores de RMSE diferentes.

O eixo horizontal de cada gráfico representa o erro quadrático médio (RMSE), enquanto o eixo vertical corresponde ao ganho de desempenho computacional em relação ao MVF em execução serial. Na Figura 6a, é destacada a região de maior interesse, posteriormente ampliada na Figura 6b, onde a arquitetura que apresentou o melhor equilíbrio entre precisão e eficiência computacional é assinalada por uma estrela vermelha. Os gráficos na parte inferior apresentam as mesmas arquiteturas, sendo codificadas por número de camadas (Figura 6d) e por número de neurônios (Figura 6c), permitindo

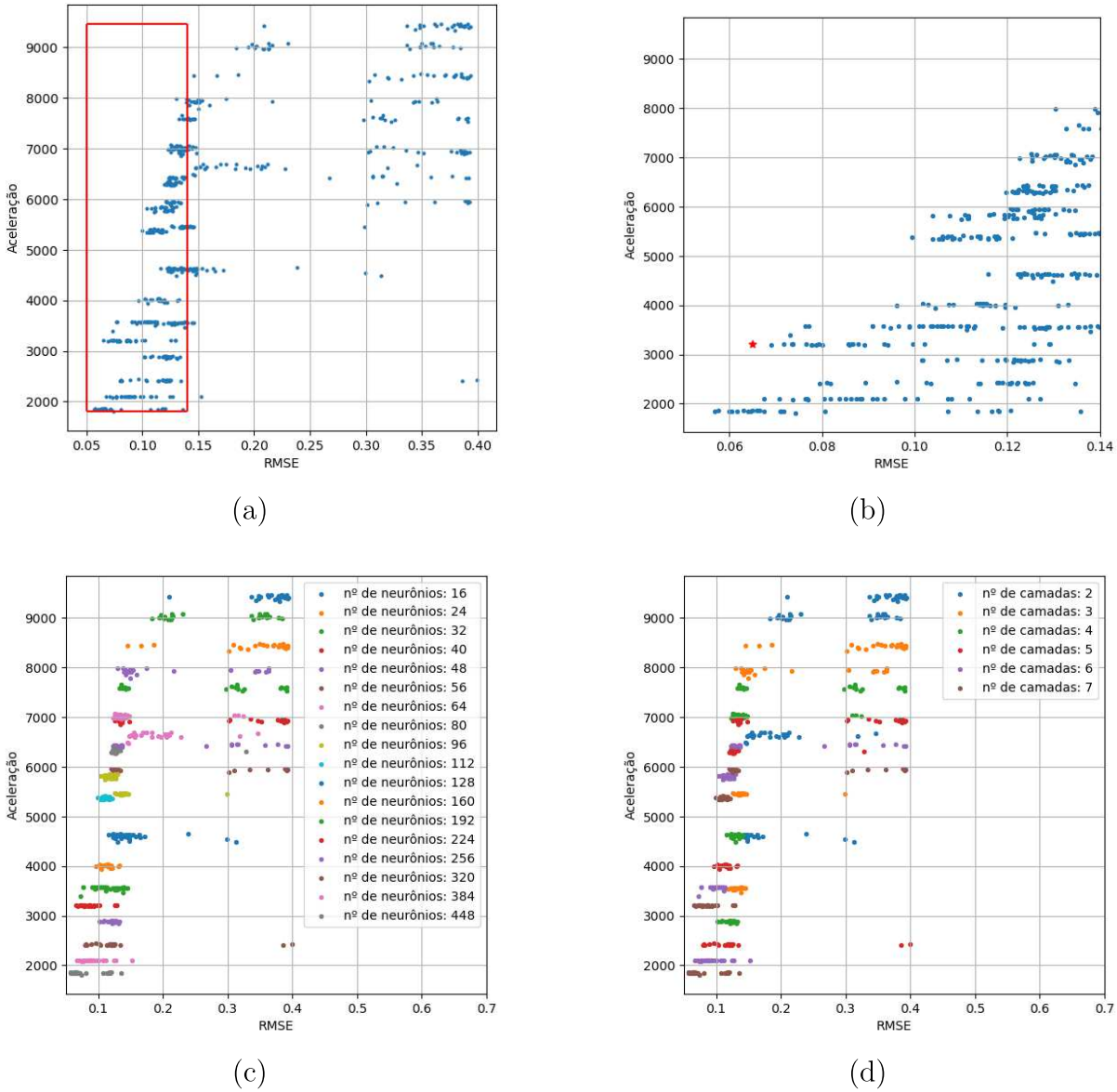


Figura 6 – Busca em grade para escolha da arquitetura.

visualizar a distribuição de desempenho em função desses dois parâmetros estruturais da rede.

A arquitetura escolhida é composta por 7 camadas ocultas contendo 32 neurônios. Em seu treino, os melhores parâmetros do otimizador ADAM foram $\beta_1 = 0{,}825$ e $\beta_2 = 0{,}99$. Essa combinação resultou no melhor compromisso em termos de acurácia e custo computacional para o problema proposto, apresentando um RMSE de 0,065, um MAE de 1,799 e uma aceleração computacional média de 3.216 ± 23, calculada em relação à execução sequencial do MVF.

4.3 COMPARAÇÃO ENTRE MODELOS

Nesta seção, são avaliadas a robustez e a eficiência de diferentes estratégias de aprendizado de máquina aplicadas à resolução de sistemas dinâmicos baseados em equações diferenciais, com foco na modelagem da dinâmica espaço-temporal de concentrações biológicas. A análise é dividida em dois eixos principais:

- a capacidade dos modelos de representar a solução latente subjacente ao sistema, mesmo em regiões não diretamente amostradas;
- a sensibilidade do desempenho em relação à quantidade de amostras utilizadas no treinamento;

No primeiro eixo, busca-se compreender em que medida os modelos são capazes de capturar com fidelidade as características qualitativas e quantitativas da solução de referência, analisando aspectos como a preservação de frentes de propagação, regiões estacionárias e padrões de ativação. No segundo, investiga-se a relação entre a densidade amostral e a acurácia das previsões, bem como o impacto dessa variação no custo computacional de treinamento. Essa abordagem integrada permite avaliar os limites de generalização das redes consideradas e orientar o uso eficiente de dados em contextos de escassez informacional.

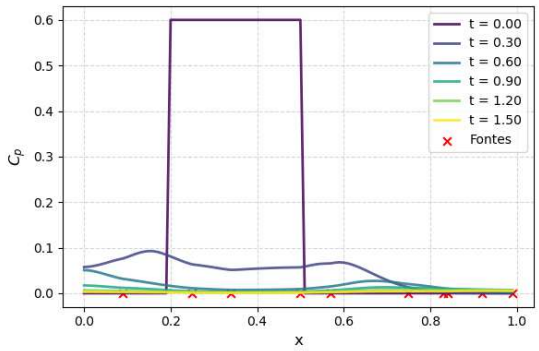
4.3.1 Capacidade de representação da solução latente

A Figura 7 apresenta a evolução espacial das concentrações de patógenos (C_p, à esquerda) e leucócitos (C_l, à direita), para diferentes instantes de tempo, simuladas com MVF, PINN e NN. Os pontos em "X" vermelhos indicam as posições das fontes de leucócitos.

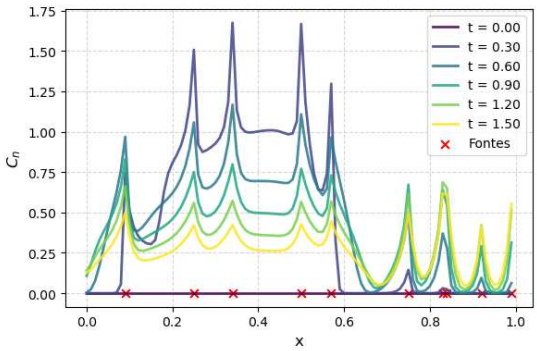
O MVF, representado na primeira linha (Figuras 7a e 7b), fornece a solução de referência adotada nesta análise. Esse referencial foi adotado por ser um modelo validado na representação de um processo inflamatório real em um paciente de miocardite infecciosa (REIS et al., 2019). A dinâmica simulada evidencia comportamentos distintos entre as duas populações modeladas. A concentração de patógenos (C_p) apresenta uma fase inicial transiente, caracterizada por difusão a partir da condição inicial, seguida por uma fase estacionária em que os níveis de C_p permanecem praticamente constantes e próximos de zero ao longo do tempo. Esse comportamento está associado à extinção dos patógenos, promovida pela ação eficiente do sistema imune.

Por outro lado, a concentração de leucócitos (C_l) exibe uma resposta rápida e espacialmente estruturada, impulsionada pelos termos de advecção-difusão que representam o fenômeno de quimiotaxia[1]. Os leucócitos são ativados em regiões próximas às fontes de

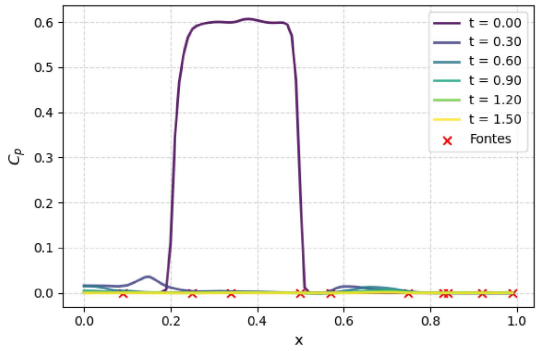
[1] Mecanismo pelo qual as células imunes se movimentam orientadas por gradientes químicos



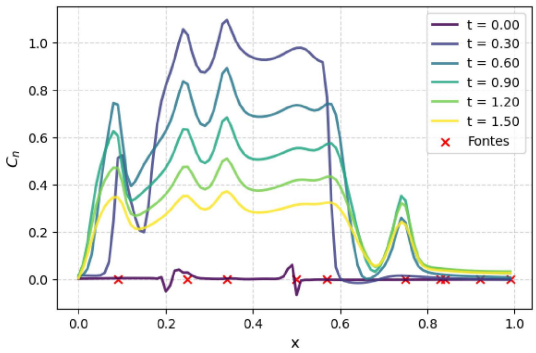
(a) MFV – concentração de patógenos (C_p).



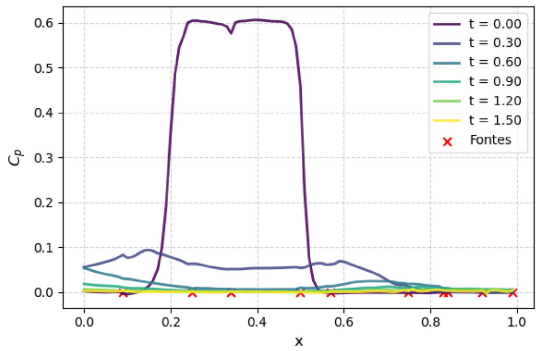
(b) MFV – concentração de leucócitos (C_l).



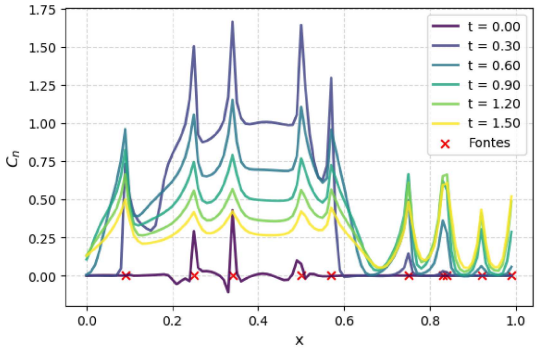
(c) PINN – concentração de patógenos (C_p).



(d) PINN – concentração de leucócitos (C_l).



(e) Rede neural clássica – concentração de patógenos (C_p).



(f) Rede neural clássica – concentração de leucócitos (C_l).

Figura 7 – Comparação temporal das curvas de concentração de patógenos (C_p) e leucócitos (C_l) geradas pelos métodos MVF, PINN e NN.

patógenos e exibem picos bem definidos ao longo do domínio, os quais evoluem de forma coordenada no tempo. A agilidade dessa resposta é condizente com a atuação do sistema imune inato, cuja função primária é fornecer uma defesa imediata frente à presença de agentes infecciosos.

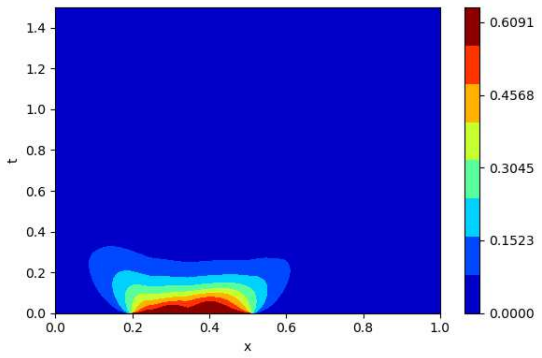
Na segunda linha (Figuras 7c e 7d), observa-se o desempenho da PINN, que apresenta limitações relevantes. Apesar de capturar qualitativamente a tendência de dissipação dos patógenos, a rede tende a reduzir os valores de C_p rapidamente, mesmo quando a solução latente se mantém praticamente estacionária. Esse comportamento sugere uma dificuldade da PINN em lidar com regiões quase invariantes no tempo, levando à subestimação das concentrações. Para a variável C_l, embora as frentes de ativação sejam reproduzidas com algum realismo, a PINN falha em capturar adequadamente a contração das concentrações de leucócitos na extremidade direita do domínio ($x \to 1$), gerando artefatos e suavizações excessivas que distorcem a dinâmica esperada.

Na terceira linha (Figuras 7e e 7f), os resultados da NN demonstram, surpreendentemente, desempenho superior em diversos aspectos. A NN consegue manter a forma da solução estacionária de C_p ao longo do tempo, além de reproduzir com maior fidelidade os picos de ativação dos leucócitos, inclusive nas regiões de maior inclinação da solução. Essa vantagem pode ser atribuída à capacidade da NN de ajustar-se diretamente aos dados, sem a imposição de restrições físicas que, neste caso específico, parecem induzir a PINN a um viés excessivo em direção à suavização. Esse fenômeno está relacionado à descontinuidade do termo fonte de leucócitos, que acaba penalizando as restrições físicas.

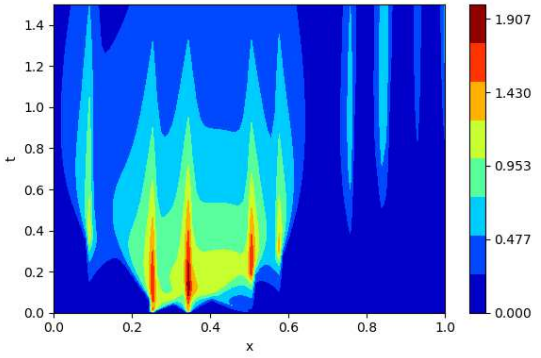
A Figura 8 fornece uma visualização espaço-temporal das concentrações de patógenos (C_p, primeira coluna) e de leucócitos (C_l, segunda coluna) ao longo do domínio espacial (x) e do tempo (t), utilizando MVF (Figuras 8b e 8a) e uma PINN, (Figuras 8d e 8c). As Figuras 8f e 8e mostram o erro absoluto entre a solução da PINN e a referência MVF. As escalas de cores representam a magnitude das variáveis em estudo ou do erro, conforme indicado nas barras laterais de cada gráfico.

No caso dos patógenos (C_p), observa-se que a PINN não é capaz de manter a solução estacionária presente na solução de referência (MVF). A dissipação precoce de C_p ao longo do tempo na solução por PINN indica uma tendência da rede em suavizar excessivamente a dinâmica, possivelmente devido à predominância de termos de dissipação na função de perda. O erro absoluto (terceira coluna, superior) revela discrepâncias localizadas próximas ao instante inicial e na região central do domínio, com valores consideráveis, o que confirma a subestimação das concentrações pela PINN.

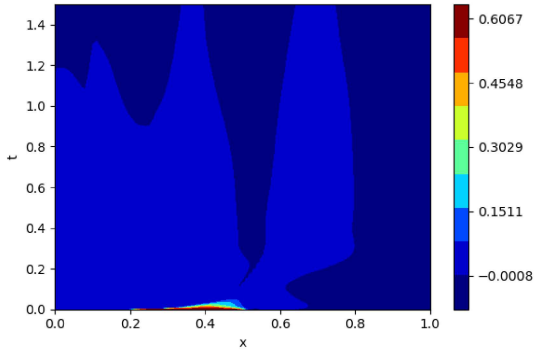
Para os leucócitos (C_l), a diferença é ainda mais expressiva. A solução do MVF apresenta padrões bem definidos de ativação e contração, com frentes de propagação que variam ao longo do tempo e se concentram em torno das fontes. A PINN, por sua vez, consegue captar parcialmente a presença dessas frentes, mas falha em manter a definição



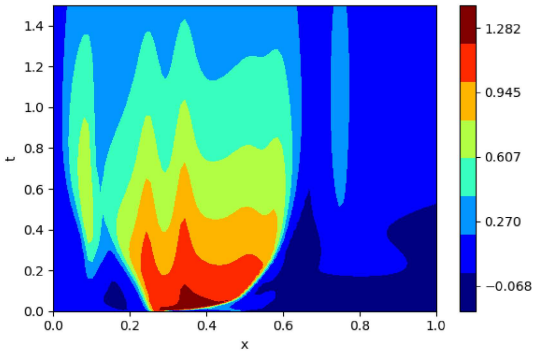
(a) Patógenos - MVF



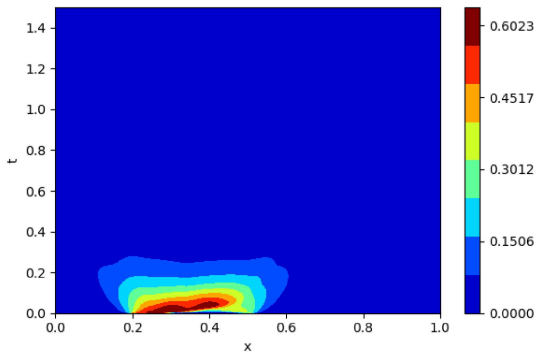
(b) Leucócitos - MVF



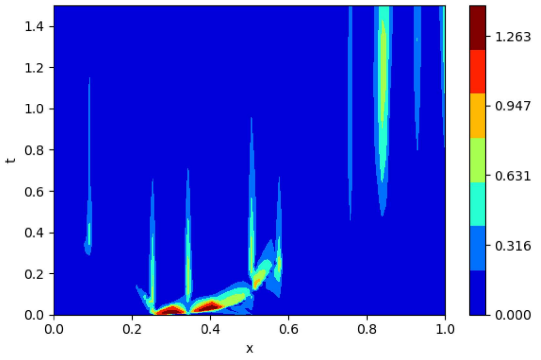
(c) Patógenos - PINN



(d) Leucócitos - PINN



(e) Erro absoluto - C_p



(f) Erro absoluto - C_l

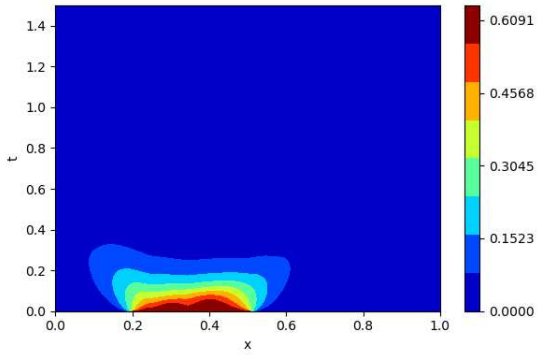
Figura 8 – Comparação das simulações via MVF e PINN para as concentrações de patógenos (C_p) e leucócitos (C_l), e os respectivos erros absolutos.

espacial e temporal nos instantes posteriores. Isso é especialmente visível na extremidade direita do domínio ($x \to 1$), onde a concentração é mal estimada. O mapa de erro absoluto (terceira coluna, inferior) evidencia essas falhas com intensas regiões de erro nos instantes intermediários, principalmente próximas às fontes e às zonas de maior variação da solução.

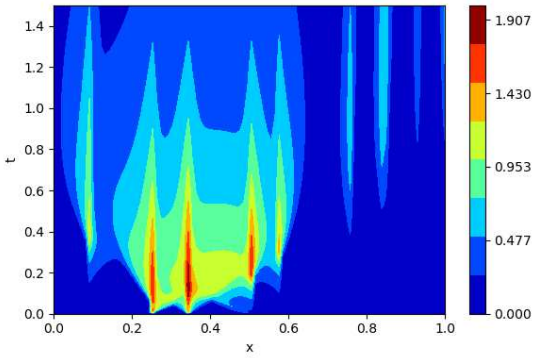
Esses resultados reforçam a limitação observada anteriormente: a PINN, embora possua uma estrutura que impõe as equações diferenciais como restrição, nem sempre consegue captar adequadamente a dinâmica de sistemas com regiões estacionárias ou com variações abruptas. A tendência à suavização e à difusão artificial pode ser explicada por

uma função de perda desbalanceada, pela normalização excessiva ou pela incapacidade da arquitetura de representar estruturas espaciais de alta frequência.

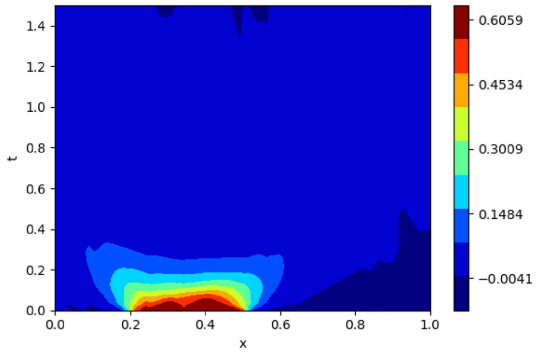
Essa visualização complementar, portanto, evidencia com clareza as regiões críticas onde a PINN apresenta maior erro e serve como ferramenta diagnóstica útil para futuras melhorias no modelo, como reponderação dinâmica dos termos da perda, ajustes de arquitetura, ou aumento da densidade de amostragem nas regiões críticas.



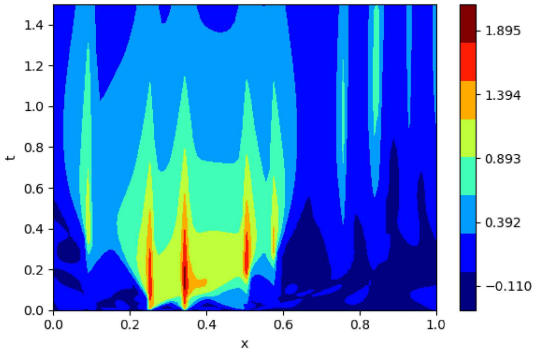
(a) Patógenos - MVF



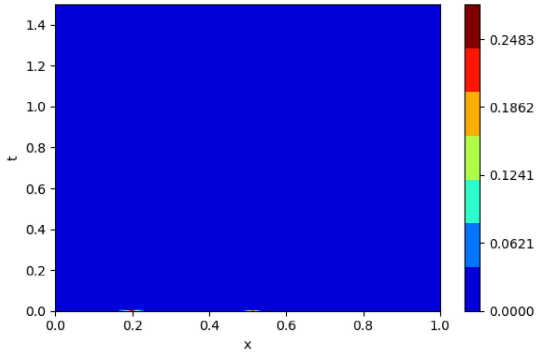
(b) Leucócitos - MVF



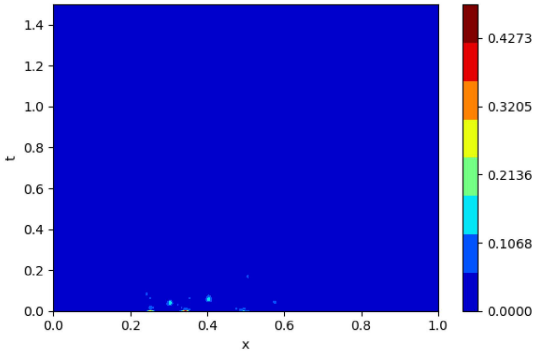
(c) Patógenos - NN



(d) Leucócitos - NN



(e) Erro absoluto - C_p



(f) Erro absoluto - C_l

Figura 9 – Comparação das simulações via MVF e NN para as concentrações de patógenos (C_p) e leucócitos (C_l), e os respectivos erros absolutos.

A Figura 9 apresenta uma comparação espaço-temporal entre a solução pelo MVF (Figuras 9a e 9a) e os resultados gerados por NN (Figuras 9a e 9a), para as concentrações

de patógenos C_p (primeira coluna) e leucócitos C_l (segunda coluna). As Figuras 9e e 9f exibem o erro absoluto entre os dois métodos. O eixo horizontal representa a coordenada espacial x, e o eixo vertical, o tempo t. As cores refletem a magnitude das concentrações ou do erro absoluto, conforme as barras de escala associadas.

Na componente C_p, observa-se que a NN é capaz de capturar adequadamente o comportamento quase estacionário da solução, com padrão de dissipação mais fiel ao MVF do que o observado na PINN (Figura 8). O erro absoluto é bastante reduzido e concentrado apenas em regiões marginais do domínio, o que indica boa capacidade de interpolação do modelo, mesmo sem incorporação explícita das equações físicas.

Já para C_l, a NN também demonstra bom desempenho, reproduzindo com fidelidade a propagação e a ativação local de leucócitos ao longo do tempo. A maioria dos picos nas posições das fontes é mantida, e a forma geral da solução é preservada. A despeito de pequenas oscilações localizadas e ruído em regiões distantes das fontes (região inferior direita), o erro absoluto é baixo em praticamente todo o domínio, o que contrasta com os erros mais intensos observados na PINN.

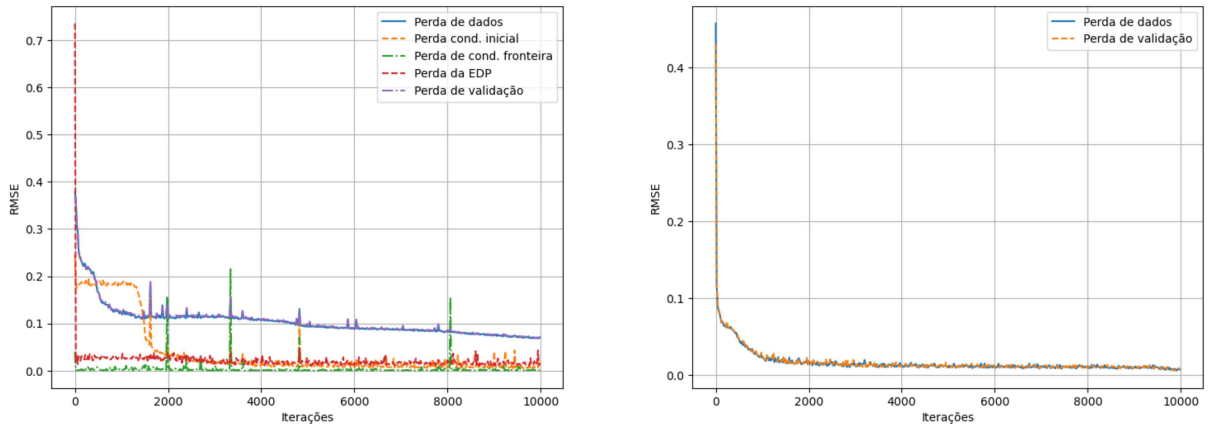
Essa análise evidencia que, apesar de não considerar diretamente as leis físicas do sistema, a rede neural convencional obteve resultados bastante precisos, especialmente para uma malha com fontes bem distribuídas e soluções relativamente suaves. Tal desempenho pode ser atribuído à maior flexibilidade da NN para ajustar-se diretamente aos dados disponíveis, o que é vantajoso em contextos com boa cobertura do espaço de entrada e ausência de ruído.

No entanto, vale destacar que esse bom desempenho pode não se manter em cenários de extrapolação, regiões de baixa densidade de dados ou quando se deseja maior interpretabilidade física. Ainda assim, neste cenário específico, a NN apresentou erro absoluto inferior à PINN e maior fidelidade à solução de referência, tanto para C_p quanto para C_l.

A Figura 10 apresenta as curvas de aprendizado obtidas durante o treinamento das arquiteturas PINN (Figura 10a) e NN (Figura 10b). Em ambas, observa-se o decaimento progressivo da perda associada aos dados e à validação, evidenciando a convergência dos modelos para soluções consistentes ao longo das iterações.

No entanto, destaca-se que a PINN exibe instabilidades pontuais em suas curvas, especialmente nas componentes da função de perda associadas às condições de fronteira e à EDP. Essas flutuações são atribuídas à descontinuidade presente na matriz booleana que representa a distribuição dos vasos sanguíneos no domínio simulado. Tais descontinuidades interferem no processo de diferenciação automática, pois esta depende da aplicação da regra da cadeia sobre funções continuamente diferenciáveis.

Sob a ótica da teoria do Neural Tangent Kernel (NTK), Wang, Yu e Perdikaris (2022) investigam as razões pelas quais PINNs podem falhar durante o treinamento. Em



(a) PINN (b) Rede neural clássica (NN)

Figura 10 – Curvas de aprendizado das PINN e NN. São apresentadas as perdas por componente (dados, condições iniciais, de fronteira, EDP e validação) ao longo das iterações.

problemas mal condicionados, o aprendizado das componentes de alta frequência da solução é dificultado, resultando em erros persistentes, mesmo quando a rede é supervisionada integralmente por restrições físicas. Uma das principais conclusões do estudo é que a estrutura degenerada do NTK em regiões com maior rigidez ou variações abruptas conduz a uma convergência enviesada, privilegiando componentes de baixa frequência. Além disso, os autores destacam que o desequilíbrio entre os termos da função de perda, especialmente entre os resíduos das equações diferenciais e as condições de contorno, é um fator determinante para o mau desempenho das PINNs. Para mitigar esse problema, propõem o ajuste dinâmico dos pesos da função de perda, estratégia que busca balancear a contribuição relativa de cada termo ao longo do treinamento, promovendo melhor representação de detalhes locais e bordas. Os resultados discutidos por Wang, Yu e Perdikaris (2022) apresentam paralelos relevantes com o presente estudo, em especial no que se refere à dificuldade de aprendizado em regiões rígidas do domínio, à persistência de erros localizados e à sensibilidade das PINNs ao balanceamento da função de perda. Observou-se que, neste trabalho, a PINN apresentou desvio sistemático nas regiões com maior descontinuidade, onde a representação da solução exata requer componentes abruptas. No entanto, diferentemente dos casos analisados por Wang, Yu e Perdikaris (2022), aqui a rigidez do domínio decorre de uma matriz booleana descontínua associada à distribuição vascular, a qual impacta diretamente os cálculos da diferenciação automática, contribuindo para a instabilidade do treinamento. Ademais, neste estudo não foi adotada a técnica de ponderação dinâmica dos termos da função de perda, o que pode ter acentuado o problema nas regiões com gradientes mais agressivos. Como discutido por Wang, Yu e Perdikaris (2022), a presença de regiões abruptas ou não suaves pode resultar em gradientes distorcidos, prejudicando o processo de retropropagação.

Esse efeito, além de dificultar a convergência suave da função de perda, pode

contribuir para os desvios observados nas regiões mais rígidas da solução, onde a PINN apresenta maior erro local. Apesar disso, nota-se que as curvas de perda associadas à validação acompanham de perto as perdas nos dados de treino para ambas as redes, indicando que os modelos convergiram sem apresentar sinais de sobreajuste ou subajuste.

Portanto, as curvas de aprendizado reforçam que, embora a PINN apresente maior sensibilidade a descontinuidades e irregularidades estruturais no domínio, ambas as abordagens são capazes de aprender adequadamente o comportamento da solução, desde que empregadas com quantidade suficiente de dados e validação adequada.

A Tabela 2 apresenta um resumo quantitativo da comparação entre os três métodos considerados neste trabalho: o método de volumes finitos (MVF) acelerado por GPU, a rede neural informada fisicamente (PINN) e a rede neural clássica (NN). São reportadas as métricas de erro (RMSE e MAE), o fator de aceleração computacional obtido em relação à execução do MVF em CPU, bem como os tempos médios de treinamento e de inferência para cada abordagem. Essa consolidação permite avaliar não apenas a acurácia das soluções, mas também o custo computacional associado à preparação e à aplicação dos modelos.

Tabela 2 – Comparação entre os métodos MVF implementado em GPU, PINN e NN, considerando as métricas de erro (RMSE e MAE), aceleração em relação ao MVF-CPU, tempo total de treinamento e tempo de inferência.

Método	**RMSE**	**MAE**	**Aceleração**	**Tempo de Treino(s)**	**Tempo de Inferência(s)**
MVF-CUDA	–	–	483 ± 8	–	0.01840 ± 0.00029
PINN	0.065	1.799	3216 ± 23	2050.09 ± 2.59	$0{,}00277 \pm 0{,}00005$
NN	0.010	0.795		267.11 ± 0.46	

Observa-se que a NN apresenta o menor RMSE quando comparada com a PINN. No entanto, o MAE da NN permanece elevado em relação ao seu RMSE, indicando a presença de erros localizados mais expressivos.

No que diz respeito ao desempenho computacional, tanto a PINN quanto a NN apresentaram acelerações superiores a 3.200 vezes, valor substancialmente superior à aceleração do MVF implementado em GPU, que atingiu aproximadamente 483 vezes. Ainda que o MVF seja computacionalmente mais simples em termos de número de operações, uma vez que realiza atualizações diretas baseadas em discretizações explícitas, o desempenho das redes neurais se mostrou superior neste cenário. Esse resultado é, à primeira vista, contraintuitivo, dado que, por exemplo, a PINN empregada possui uma arquitetura composta por 7 camadas com 32 neurônios totalmente conectados, o que implica em 32^7 operações não lineares por predição.

Entretanto, a principal justificativa para essa vantagem está na forma como o problema é paralelizado. As redes neurais não dependem da evolução temporal explícita

para computar a solução, ou seja, o domínio do tempo é tratado como uma variável de entrada e pode ser avaliado de forma totalmente paralela, assim como o espaço. Já o MVF implementado de forma explícita apresenta dependência sequencial no tempo, o que limita sua paralelização ao domínio espacial. Isso faz com que, mesmo sendo computacionalmente mais leve em termos de complexidade algébrica, o MVF sofra restrições estruturais que impedem ganhos mais expressivos de desempenho na GPU.

Apesar da eficiência na inferência, é importante observar que o tempo total de treinamento das redes neurais representa um custo relevante. No presente estudo, o tempo médio de treinamento da PINN foi de aproximadamente 2.050 segundos, enquanto a NN foi treinada em cerca de 267 segundos. Esses valores contrastam com o tempo necessário para realizar a simulação direta com MVF, indicando que, do ponto de vista do custo total, a aplicação das redes pode não ser vantajosa quando o objetivo é obter a solução para um único conjunto de condições. Contudo, esse custo de treinamento pode ser amortizado em contextos onde se deseja generalizar a solução para diferentes cenários paramétricos ou realizar estudos de sensibilidade, nos quais as redes treinadas permitem inferência direta e eficiente para diferentes pontos no espaço-tempo, sem necessidade de simulações sucessivas.

Por fim, é importante destacar que, mesmo ao considerar a versão paralelizada do MVF na GPU, cuja aceleração média em relação à execução sequencial foi da ordem de 483 ± 8 vezes, as redes neurais mantiveram vantagem substancial. Em particular, a rede neural clássica (NN) apresentou um fator de aceleração de 3.216 ± 23 em relação ao MVF em CPU, o que representa aproximadamente $6,66$ vezes mais do que o MVF mesmo após paralelização. Esse comportamento ressalta que, para problemas com forte dependência temporal, as abordagens baseadas em redes neurais oferecem não apenas boa acurácia (quando bem ajustadas), mas também um elevado potencial de aceleração da inferência, particularmente vantajoso em simulações em larga escala ou com múltiplas condições iniciais, onde o custo marginal da inferência se torna crítico.

Com base nas análises apresentadas, observa-se que diferentes arquiteturas e metodologias de treinamento possuem limitações e pontos fortes distintos no que diz respeito à fidelidade das soluções espaço-temporais obtidas. Entretanto, a qualidade das previsões também depende diretamente da quantidade de informações disponíveis no conjunto de treinamento. Diante disso, o próximo tópico busca investigar como a densidade amostral influencia a acurácia e o custo computacional do treino dos modelos, fornecendo subsídios quantitativos para o dimensionamento ótimo de dados em aplicações práticas.

4.3.2 Análise de sensibilidade à redução de amostras temporais

Nesta seção, analisamos o impacto da redução do número de amostras temporais no desempenho das abordagens PINN e NN. A motivação para essa análise reside na

necessidade de avaliar a robustez e a eficiência dos métodos quando submetidos a cenários com menor disponibilidade de dados, uma situação comum em aplicações que envolvem a modelagem de sistemas biológicos. Para tal, variamos o número de passos de tempo disponíveis para o treinamento das redes e comparamos RMSE e o tempo de execução médio obtidos em cada cenário. Os resultados apresentados a seguir permitem comparar não apenas a acurácia das soluções, mas também a relação custo-benefício de cada abordagem diante da escassez de dados temporais.

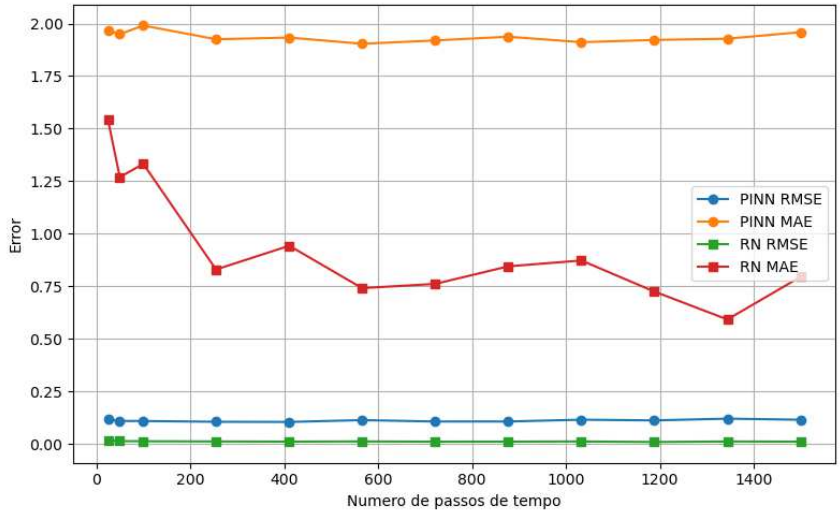


Figura 11 – Erros RMSE e MAE em função do número de passos temporais, com cobertura espacial completa, utilizados no treinamento das redes PINN e NN.

A Figura 11 apresenta a evolução dos erros RMSE e MAE para as abordagens PINN e NN, em função da quantidade de passos de tempo utilizados na base de treinamento. Cabe destacar que, a cada passo temporal, foram considerados todos os pontos disponíveis na discretização espacial do domínio, composta por 100 pontos uniformemente espaçados. Dessa forma, cada passo de tempo corresponde a 100 amostras espaço-temporais, de modo que, por exemplo, uma base com 25 passos temporais representa um total de 2.500 amostras distintas. Essa estratégia assegura uma cobertura espacial completa a cada instante avaliado, permitindo avaliar com maior precisão o impacto da redução da resolução temporal sobre o desempenho das redes.

Observa-se que os erros da PINN se mantêm altos em comparação à NN para todos os números de amostras considerados. Além disso, tanto o RMSE quanto o MAE da PINN flutuam de maneira sutil com o aumento das amostras temporais, com uma leve piora a partir de 877 passos de tempo. Tal comportamento pode indicar sensibilidade à escolha dos dados de treinamento, bem como possíveis efeitos de sobreajuste ou limitações numéricas no processo de otimização.

Para a rede neural clássica, a redução inicial no número de amostras tem impacto significativo na melhoria dos erros. No entanto, os ganhos se estabilizam a partir de aproximadamente 566 amostras, ponto em que tanto o RMSE quanto o MAE deixam de

apresentar variações expressivas. Este valor representa, portanto, um ponto de saturação, além do qual o acréscimo de amostras temporais não contribui de forma relevante para o desempenho do modelo.

A comparação entre os erros RMSE e MAE também evidencia que o MAE é sistematicamente maior, especialmente na PINN, refletindo a presença de desvios absolutos relevantes mesmo quando o erro quadrático se mantém relativamente controlado. Ademais, temos que o MAE se mostra mais sensível ao aumento de amostras de treino que o RMSE.

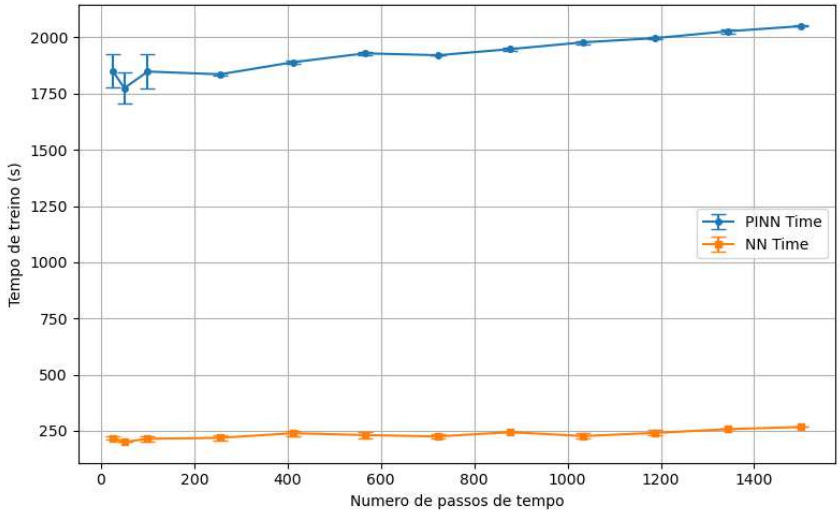
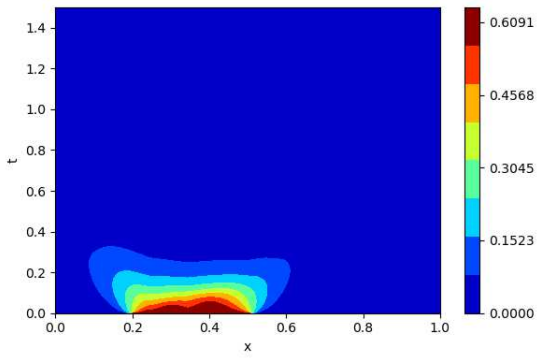


Figura 12 – Tempo de execução médio em função do número de passos temporais, com cobertura espacial completa, utilizados no treinamento das redes PINN e NN.

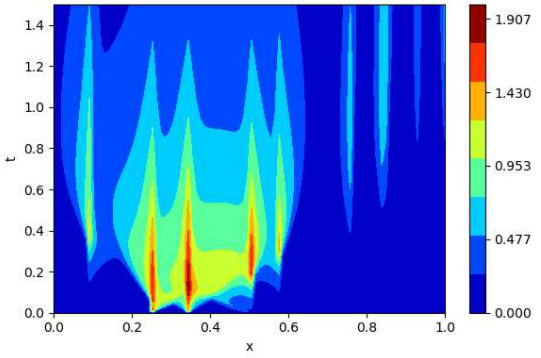
A Figura 12 apresenta os tempos médios de execução para os diferentes cenários de amostragem temporal. Nota-se que a PINN apresenta uma redução expressiva no tempo de execução conforme o número de amostras diminui, evidenciando que grande parte do custo computacional se deve ao cálculo dos termos derivados e não ao volume de dados em si. Para a NN, por outro lado, a variação no número de amostras afeta apenas marginalmente o tempo total, pois seu custo está diretamente atrelado à quantidade de dados disponíveis. Entretanto, como ambos os métodos foram treinados em paralelo, essa baixa variação pode ser explicada por uma subutilização da GPU.

A Figura 13 apresenta a comparação entre os resultados obtidos pelo MVF e aqueles obtidos por uma NN treinada com os dados gerados por MVF. As simulações também foram realizadas com 25 passos de tempo durante o treino.

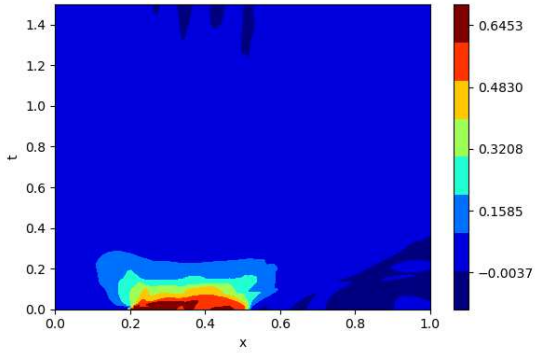
As Figuras 13c e 13d exibem as soluções de referência obtidas por MVF, enquanto as Figuras 13c e 13d mostram as soluções previstas por uma rede neural comum. Observa-se que a rede neural consegue representar de forma mais precisa os padrões gerais da evolução espaço-temporal. Entretanto, há uma perda evidente de precisão em regiões com maior variação espacial ou temporal, como mostrado nos mapas de erro das Figuras 13e e 13f. Além disso, é importante destacar que ela também tem erros mais proeminentes nos



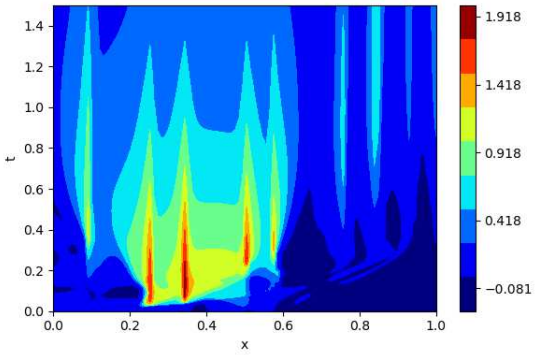
(a) Patógenos - MVF



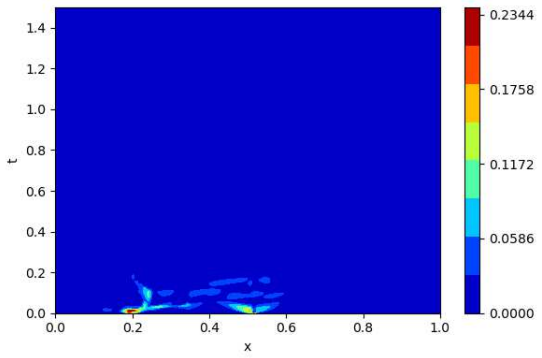
(b) Leucócitos - MVF



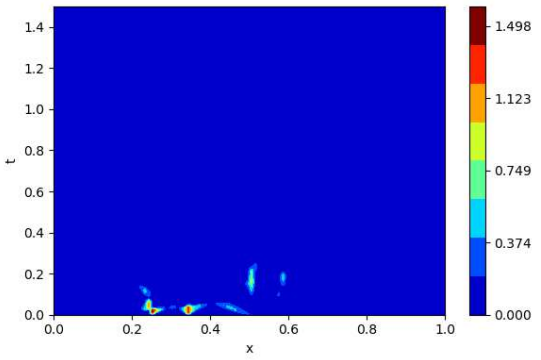
(c) Patógenos - NN



(d) Leucócitos - NN



(e) Erro absoluto - C_p



(f) Erro absoluto - C_l

Figura 13 – Mapas de calor representando a concentração de patógenos (C_p) e leucócitos (C_l) simuladas via MVF e rede neural padrão (NN) utilizando 25 pontos no tempo, bem como o erro absoluto associado às soluções por NN.

instantes iniciais da simulação.

Quando comparado ao modelo treinado com o conjunto completo de amostras temporais, o resultado com 566 amostras mantém boa fidelidade qualitativa, sendo capaz de representar com acurácia os principais aspectos da dinâmica da concentração de patógenos. No entanto, pequenas diferenças de amplitude e regularidade ainda podem ser notadas, particularmente nas regiões próximas às fontes.

Este comportamento evidencia que o uso de 566 amostras representa um ponto

intermediário eficaz entre custo computacional e qualidade da solução. Por outro lado, o modelo treinado com apenas 25 amostras apresentou perdas expressivas de precisão, com maiores desvios da solução de referência e tendência a suavizar excessivamente regiões de transição abrupta. Tais limitações são coerentes com a escassez de informação temporal fornecida durante o treinamento, comprometendo a generalização do modelo.

A Figura 14 exibe a comparação entre a solução de referência obtida por meio do MVF e os resultados gerados por uma NN treinada com dados provenientes do MVF. O treinamento da rede foi realizado utilizando 566 passos temporais.

As Figuras 14c e 14d ilustram a evolução espaço-temporal das concentrações de patógenos (C_p) e leucócitos (C_l), respectivamente, obtidas a partir da rede neural clássica treinada com 566 amostras temporais. Já as Figuras 14e e 14f exibem os mapas de erro absoluto correspondentes, com relação à solução de referência calculada via MVF.

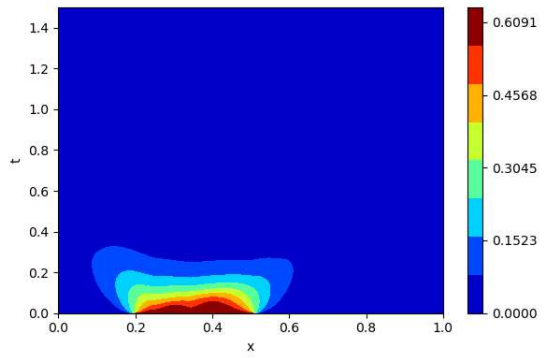
Comparando com os resultados obtidos com apenas 25 amostras temporais, observa-se uma melhora substancial na qualidade das predições. Os mapas previstos demonstram maior coerência entre os instantes temporais e maior aderência às regiões onde ocorrem fenômenos relevantes, como difusão e advecção. Isso é particularmente evidente nos perfis de C_l, nos quais a estrutura das fontes é mais bem delineada, embora ainda existam picos não físicos e padrões oscilatórios residuais, principalmente nos intentes inciais da simulação.

Ainda, os resultados com 566 amostras aproximam-se fortemente daqueles obtidos com o conjunto completo de dados temporais, tanto em termos qualitativos quanto quantitativos. A estrutura das soluções é preservada, as principais regiões de concentração são bem capturadas, e os erros residuais são pontuais e de baixa intensidade. Isso indica que 566 amostras temporais representam um ponto de equilíbrio eficiente para o treinamento da NN, fornecendo resultados comparáveis ao cenário de máxima informação, porém com um custo computacional consideravelmente reduzido.

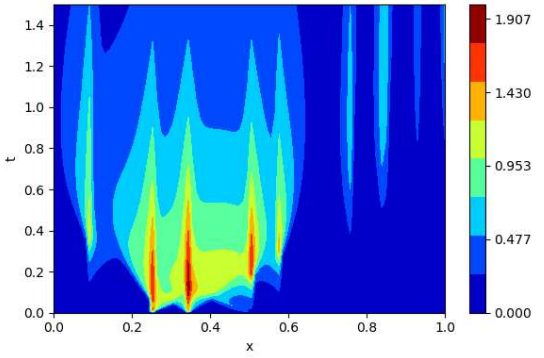
Essa constatação reforça a ideia de que, para redes neurais clássicas, há um limiar de saturação a partir do qual o aumento no número de amostras não implica melhorias significativas na solução. Assim, o uso de 566 amostras mostra-se uma escolha vantajosa sob a perspectiva de desempenho versus custo computacional.

Os resultados apresentados ao longo desta seção evidenciam diferenças significativas na sensibilidade ao número de amostras temporais entre os métodos analisados. Em particular, observou-se que as NN apresentam forte dependência da quantidade de dados para alcançar soluções estáveis.

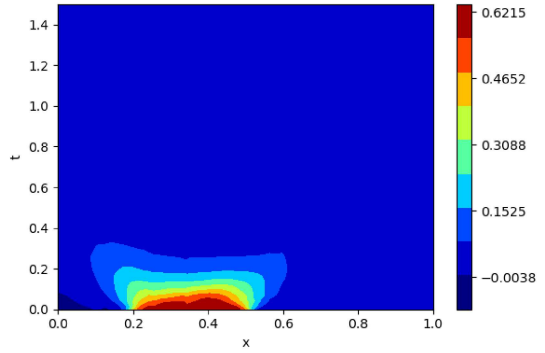
Entretanto, ao considerar o custo de treinamento, verifica-se que o tempo requerido para treinar tanto a PINN quanto a NN não se justifica quando comparado ao tempo necessário para simular diretamente com o MVF. Essa constatação é especialmente relevante em contextos onde o objetivo é a obtenção pontual da solução, uma vez que o MVF é



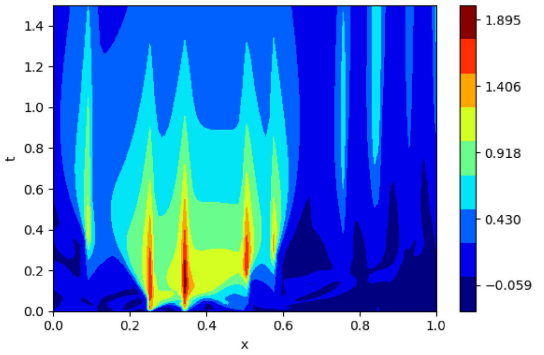
(a) Patógenos - MVF



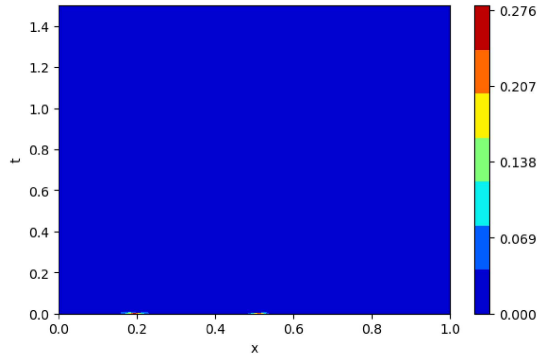
(b) Leucócitos - MVF



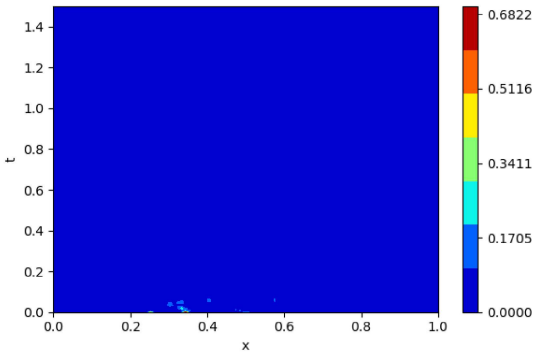
(c) Patógenos - NN



(d) Leucócitos - NN



(e) Erro absoluto - C_p



(f) Erro absoluto - C_l

Figura 14 – Mapas de calor representando a concentração de patógenos (C_p) e leucócitos (C_l) simuladas via MVF e rede neural padrão (NN), bem como o erro absoluto associado às soluções por NN.

computacionalmente mais leve e requer menos operações. A PINN utilizada neste estudo, por exemplo, possui 7 camadas com 32 neurônios totalmente conectados, resultando em milhares de conexões não lineares a serem avaliadas a cada predição.

Contudo, vale destacar que esse custo de treinamento pode ser diluído em aplicações que exigem generalização paramétrica. Nesses casos, uma vez treinada, a rede pode ser utilizada para prever múltiplos cenários, variando-se parâmetros do modelo sem necessidade de serem treinadas novamente. Essa característica é particularmente vantajosa em estudos

de sensibilidade ou em tarefas de inferência inversa, onde seria necessário resolver o sistema para inúmeras configurações. Além disso, como as redes neurais operam tratando o tempo como uma variável de entrada, sua inferência pode ser feita para qualquer instante diretamente, ao passo que o MVF explícito requer a integração sequencial no tempo. Assim, embora o custo de treinamento inicial das redes seja elevado, o custo marginal de novas inferências é substancialmente inferior, tornando-as atrativas em aplicações que envolvem muitas simulações subsequentes.

5 CONCLUSÃO

Com base nos objetivos delineados e nas contribuições alcançadas, esta conclusão retoma os principais propósitos que nortearam o desenvolvimento deste trabalho. O estudo teve como foco central investigar o potencial de DNNs na modelagem da resposta imunológica local em quadros de miocardite infecciosa, com ênfase na representação espaço-temporal das concentrações de patógenos e leucócitos. Para tanto, foram implementadas e analisadas formulações baseadas tanto em PINNs quanto em NNs, tendo como referência um modelo matemático previamente validado (REIS et al., 2019) e resolvido pelo MVF. A comparação entre essas abordagens buscou não apenas avaliar a acurácia e o custo computacional, mas também examinar a sensibilidade ao número de amostras de treinamento e à arquitetura adotada. A partir dessa perspectiva, a conclusão sintetiza os principais achados e limitações observadas ao longo dos experimentos, destacando as implicações práticas e metodológicas da aplicação de PINNs e NNs em contextos de imunologia computacional.

A comparação entre as abordagens PINN e NN permitiu uma análise detalhada de seus respectivos desempenhos frente à tarefa de modelagem da dinâmica espaço-temporal de concentrações biológicas. Ambas as técnicas foram avaliadas considerando a acurácia das soluções obtidas, a robustez frente à variação da densidade amostral e o custo computacional associado tanto ao processo de treinamento quanto à inferência.

As PINNs mostraram-se menos sensíveis à redução do número de amostras temporais utilizadas durante o treinamento, o que evidencia a contribuição da incorporação explícita de restrições físicas como mecanismo de regularização. Essa característica confere ao modelo uma estrutura mais rígida, capaz de manter coerência global na solução mesmo sob escassez de dados, ainda que sem garantir alta acurácia. Observou-se, entretanto, que essa rigidez compromete a capacidade de representar adequadamente fenômenos localizados ou de rápida variação, resultando em suavização excessiva das soluções e subestimação de concentrações em regiões estacionárias ou com gradientes acentuados.

Por outro lado, as NNs destacaram-se pela capacidade de representar com elevada fidelidade os perfis espaço-temporais das variáveis modeladas, mesmo em regiões de elevada complexidade. A NN foi capaz de manter a forma da solução de referência em instantes avançados da simulação, preservando tanto o comportamento estacionário dos patógenos quanto os picos de ativação de leucócitos. Tal desempenho, no entanto, mostrou-se fortemente dependente da densidade amostral, com degradação significativa da solução em cenários com número muito reduzido de amostras temporais.

No que tange à aceleração, ambas as abordagens baseadas em redes neurais superaram substancialmente o desempenho do MVF, mesmo em sua versão paralelizada via GPU. Essa superioridade decorre da possibilidade de avaliação paralela dos domínios espaço-temporais, uma vez que as redes tratam tempo e espaço como variáveis de entrada

independentes, sem necessidade de integração sequencial no tempo.

Considerando o conjunto dos resultados obtidos, observa-se que, para o problema analisado, a rede neural clássica demonstrou desempenho superior. A PINN, mostrou limitações na representação de dinâmicas localizadas e na preservação de estruturas estacionárias.

Essas limitações estão relacionadas com a descontinuidade presente na matriz booleana que representa a distribuição dos vasos sanguíneos no domínio simulado. Tais descontinuidades interferem diretamente na diferenciação automática, uma vez que este processo depende da aplicação da regra da cadeia sobre funções suavemente diferenciáveis. A presença de regiões abruptas ou não suaves pode distorcer os gradientes calculados, afetando a retropropagação e comprometendo a convergência da rede. Esse fenômeno pode explicar, ao menos em parte, os desvios de inferência observados nas regiões com maior complexidade geométrica ou comportamento não contínuo da solução.

Ainda assim, ao se considerar o custo computacional total, compreendendo tempo de treinamento e inferência, verifica-se que as redes neurais, não são vantajosas quando o objetivo é a simples obtenção pontual da solução, já que o tempo necessário para treinar o modelo excede, com folga, o tempo de simulação direta via MVF. Esse resultado é particularmente relevante dado que o MVF requer um número muito menor de operações para calcular cada ponto da solução, enquanto redes como a PINN, com sete camadas totalmente conectadas, demandam milhares de operações por predição.

Contudo, tal custo de treinamento pode ser amortizado em aplicações que envolvem múltiplas inferências ou variações paramétricas. Nesses contextos, uma rede previamente treinada pode ser generalizada para diferentes configurações, permitindo previsões rápidas sem a necessidade serem treinadas novamente. Além disso, por tratarem tempo e espaço como entradas, as redes são capazes de realizar inferência diretamente em qualquer instante desejado, ao passo que o MVF explícito precisa integrar sequencialmente no tempo. Essa diferença estrutural viabiliza ganhos significativos em contextos como estudos de sensibilidade, simulações em larga escala ou problemas inversos, nos quais o custo marginal da inferência via redes é consideravelmente inferior ao custo acumulado da resolução sequencial via métodos tradicionais.

Dessa forma, conclui-se que, no contexto específico considerado neste estudo, a rede neural clássica representa a escolha mais eficaz, proporcionando o melhor compromisso entre precisão, eficiência computacional e fidelidade às soluções de referência. Ainda assim, ressalta-se o potencial das PINNs em aplicações com menor densidade de dados e maior exigência de consistência física, o que as torna uma alternativa promissora em diferentes contextos de modelagem científica.

5.1 LIMITAÇÕES DO ESTUDO

Apesar dos resultados promissores obtidos ao longo deste trabalho, é importante reconhecer algumas limitações que podem ter influenciado a generalização dos achados e que devem ser consideradas em estudos futuros.

Em primeiro lugar, todas as redes neurais avaliadas, tanto a PINN quanto a NN, foram treinadas utilizando exclusivamente a função de ativação *tanh*. Ainda que essa escolha tenha sido justificada com base na literatura e em sua suavidade, a limitação à avaliação de uma única função de ativação pode ter restringido o desempenho das arquiteturas, sobretudo em regiões com comportamento abrupto. A inclusão de outras funções, como *ReLU*, *Swish* ou *GELU*, poderia revelar arquiteturas mais adequadas ao problema modelado.

Adicionalmente, a aplicação foi restrita a um modelo unidimensional, com dinâmica espaço-temporal representada em malhas regulares. Embora essa abordagem tenha permitido análises controladas e reprodutíveis, ela não explora integralmente o potencial de paralelização e otimização da GPU em métodos clássicos como o MVF, especialmente em contextos bidimensionais ou tridimensionais. Situações com maior dimensionalidade espacial podem representar cenários mais desafiadores e realistas para a avaliação do desempenho relativo entre os métodos.

Outro ponto a ser considerado refere-se à utilização de dados sintéticos gerados a partir de uma solução de referência conhecida. Embora tal escolha seja válida para validação inicial de modelos, sua aplicação em dados reais — que frequentemente apresentam ruídos, incertezas e desbalanceamentos — pode demandar estratégias adicionais de regularização e validação cruzada, além de modificações na formulação da função de perda.

Um aspecto relevante que não foi explorado neste estudo diz respeito à robustez dos modelos frente à presença de ruído nos dados de entrada. Todos os experimentos foram conduzidos com dados sintéticos gerados diretamente a partir da solução de referência obtida via MVF, sem adição de perturbações aleatórias que simulem incertezas típicas de medições experimentais. Em contextos reais, no entanto, é comum que os dados disponíveis estejam sujeitos a ruído, variabilidade amostral e imprecisões de natureza instrumental ou biológica. A avaliação da resiliência das abordagens consideradas (em especial as PINNs e NNs) em cenários com dados ruidosos representa, portanto, uma extensão importante e necessária para confirmar a aplicabilidade prática dos modelos. Estudos anteriores indicam que a incorporação de conhecimento físico tende a favorecer a estabilidade em tais contextos, mas a validação empírica dessa hipótese, no escopo da resposta imunológica modelada, permanece em aberto.

Outro fator limitante diz respeito à utilização de pesos fixos na composição da função de perda durante o treinamento das PINNs. Conforme demonstrado na literatura,

a escolha inadequada dos pesos relativos entre os diferentes termos, como os resíduos das EDPs, as condições de contorno e os dados observacionais, pode resultar em um treinamento desequilibrado, no qual certas regiões do domínio são priorizadas em detrimento de outras, especialmente em problemas mal condicionados. Estratégias de ponderação dinâmica, como o reescalonamento adaptativo baseado na norma dos gradientes ou no erro relativo de cada termo, têm se mostrado eficazes para mitigar esse problema, promovendo um aprendizado mais uniforme e favorecendo a representação de componentes de alta frequência. No presente trabalho, optou-se pela adoção de pesos constantes, definidos empiricamente, tanto por simplicidade quanto por foco comparativo entre abordagens. No entanto, reconhece-se que a ausência de um mecanismo adaptativo de balanceamento pode ter contribuído para os erros localizados observados nas regiões com gradientes mais acentuados, e representa uma oportunidade clara de aprimoramento em estudos futuros.

Além disso, ressalta-se que os resultados apresentados nos gráficos e mapas de calor foram obtidos a partir dos mesmos pontos utilizados no cálculo da perda de dados. Dessa forma, as métricas de erro reportadas refletem a capacidade das redes em replicar a solução de referência dentro do mesmo domínio de treinamento, mas não fornecem uma estimativa clara da generalização dos modelos para novos pontos espaço-temporais ou para condições iniciais e de contorno distintas. Avaliações futuras deverão incluir testes de extrapolação, particionamento dos dados e métricas de desempenho sobre domínios fora da amostra para garantir uma análise mais abrangente da capacidade preditiva das redes.

Outra limitação metodológica refere-se à desconsideração do tempo de compilação inicial das funções decoradas com *@cuda.jit* durante a avaliação do desempenho computacional. Ao utilizar a biblioteca Numba, a compilação JIT (*Just-In-Time*) dos *kernels* CUDA ocorre na primeira execução de cada função, introduzindo um *overhead* que não foi contabilizado nos testes de aceleração. Essa decisão foi motivada por dois fatores. Primeiro, esse tempo adicional é uma particularidade da implementação em Python com Numba e não está presente em implementações equivalentes escritas diretamente em CUDA C/C++, que representam o estado da arte em desempenho. Segundo, em cenários reais de aplicação, nos quais os *kernels* compilados são reutilizados em múltiplas simulações sucessivas, esse tempo inicial torna-se desprezível frente ao tempo total de execução, sendo rapidamente diluído ao longo das execuções. Assim, a métrica de aceleração reportada busca refletir o desempenho prático efetivo da simulação após a etapa de compilação, tal como se espera em aplicações reais de maior escala.

Por fim, destaca-se uma limitação metodológica associada à escolha da topologia de rede empregada na comparação entre as abordagens PINN e NN. Ambas as arquiteturas foram avaliadas utilizando a mesma configuração de camadas e neurônios, sem a realização de uma busca independente e específica de topologias para cada tipo de rede. Essa decisão visou isolar o efeito da imposição explícita de restrições físicas sobre o desempenho do modelo, mantendo constante a complexidade estrutural das redes. No entanto, essa

simplificação pode ter introduzido um viés na comparação, uma vez que a arquitetura ideal para uma PINN pode não coincidir com a mais adequada para uma NN tradicional. Abordagens mais rigorosas deveriam empregar estratégias de otimização de arquitetura, como busca em grade ou algoritmos evolutivos, aplicadas separadamente a cada classe de rede, de modo a garantir uma comparação justa entre seus desempenhos máximos potenciais. A investigação desse aspecto constitui uma linha promissora para trabalhos futuros.

Dessa forma, reconhece-se que os resultados aqui apresentados devem ser interpretados à luz dessas limitações. Estudos futuros poderão explorar alternativas de ativação, domínios de maior complexidade, dados experimentais e métricas mais especializadas para ampliar a aplicabilidade e a robustez das conclusões obtidas.

5.2 TRABALHOS FUTUROS

A partir das limitações identificadas e dos resultados obtidos, diversas direções podem ser exploradas em trabalhos futuros com o objetivo de ampliar a robustez e a aplicabilidade das abordagens baseadas em redes neurais para resolução de sistemas dinâmicos.

Uma das principais frentes de investigação consiste na exploração de estratégias mais eficazes de ponderação dos termos da função de perda em PINNs. Conforme discutido no Capítulo 1, a má formulação das penalidades nos termos residuais pode comprometer a convergência do treinamento, sobretudo em domínios com múltiplas escalas. Os autores propõem uma reponderação informada pela estrutura do *Neural Tangent Kernel* (NTK), capaz de mitigar o desequilíbrio entre os diferentes componentes da função de perda e melhorar a propagação de gradientes nas regiões críticas da solução. A implementação de estratégias adaptativas com base nessa abordagem pode representar um avanço significativo na estabilidade e na acurácia das PINNs.

Além disso, a seleção de arquiteturas adequadas permanece um dos principais desafios na aplicação de DNNs. Embora este trabalho tenha utilizado a técnica de busca em grade, métodos mais sofisticados de otimização estrutural, como algoritmos genéticos, podem ser empregados para explorar de maneira mais eficiente o espaço de configurações possíveis. Esses algoritmos evolutivos têm se mostrado promissores na identificação de arquiteturas com melhor capacidade de generalização e menor custo computacional. Nesse sentido, pretende-se investigar a aplicabilidade de métodos baseados em *Neuroevolution*, tanto para PINNs quanto para redes convencionais, considerando critérios múltiplos de desempenho.

Como proposta para trabalhos futuros, também se destaca a avaliação da capacidade de generalização das NNs e PINNs no aprendizado de superfícies paramétricas de solução. O objetivo é investigar se tais modelos são capazes de estimar adequadamente a resposta

inflamatória para diferentes valores de parâmetros fisiológicos, como os coeficientes de difusão, quimiotaxia e fagocitose, sem a necessidade de novo treinamento a cada variação. Para isso, pretende-se incluir tais parâmetros como entradas adicionais na rede, de modo que a solução aprendida represente uma família contínua de respostas. Além disso, a análise será estendida à capacidade das redes em inferir corretamente a solução em pontos espaço-temporais não utilizados durante o treinamento, o que permitirá uma avaliação mais rigorosa da sua habilidade de generalização.

Por fim, outra linha de continuidade natural deste trabalho reside na extensão dos modelos para domínios bidimensionais e tridimensionais. Tais casos não apenas ampliam a relevância prática dos modelos em aplicações reais da engenharia biomédica, como também desafiam os métodos propostos quanto à escalabilidade computacional e capacidade de representação espacial.

REFERÊNCIAS

ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, v. 6, p. 52138–52160, 2018.

AMARI, S.-i. Backpropagation and stochastic gradient descent method. *Neurocomputing*, Elsevier, v. 5, n. 3, p. 185–196, 1993.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, v. 5, n. 2, p. 157–166, 1994.

BOTTOU, L. Large-scale machine learning with stochastic gradient descent. In: SPRINGER. *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. [S.l.], 2010. p. 177–186.

CAFORIO, A. L. et al. Current state of knowledge on aetiology, diagnosis, management, and therapy of myocarditis: a position statement of the european society of cardiology working group on myocardial and pericardial diseases. *European heart journal*, Oxford University Press, v. 34, n. 33, p. 2636–2648, 2013.

CHEN, H.-Y. et al. Physics-informed graph neural network for predicting fluid flow in porous media. *Petroleum Science*, 2025. ISSN 1995-8226. Disponível em: <https://www.sciencedirect.com/science/article/pii/S199582262500216X>.

CYBENKO, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, v. 2, n. 4, p. 303–314, 1989.

DUCHI, J.; HAZAN, E.; SINGER, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, v. 12, n. 7, 2011.

FERNANDES, T. E. *2D_imune_edema_pinn: Repositório de códigos-fonte*. 2025. <https://github.com/Esterci/2D_imune_edema_pinn>. Acesso em: 10 jul. 2025.

FERNANDES, T. E. et al. Bridging accuracy and efficiency: The role of pinns in immune response simulations. *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 6612–6619, 2024.

GOODFELLOW, I. et al. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1.

HUMFELD, K. D. et al. Co-training of multiple neural networks for simultaneous optimization and training of physics-informed neural networks for composite curing. *Composites Part A: Applied Science and Manufacturing*, v. 193, p. 108820, 2025. ISSN 1359-835X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1359835X25001149>.

KINGMA, D. P.; BA, J. L. Adam: A method for stochastic optimization. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. San Diego, CA, USA: [s.n.], 2015. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. Disponível em: <https://arxiv.org/abs/1412.6980>.

LAM, S. K. et al. *Numba: a Lexically Scoped JIT Compiler in Python*. 2024. <https://numba.pydata.org/>. Acesso em: 10 jul. 2025.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, 2015. Nature Publishing Group.

LEVEQUE, R. J. *Finite Difference Methods for Ordinary and Partial Differential Equations*. [S.l.]: Society for Industrial and Applied Mathematics, 2007.

LOURENÇO, W. d. J. et al. A poroelastic approach for modelling myocardial oedema in acute myocarditis. *Frontiers in Physiology*, Frontiers Media SA, v. 13, p. 888515, 2022.

MARTIN, C. H. et al. Ep-pinns: Cardiac electrophysiology characterisation using physics-informed neural networks. *Frontiers in Cardiovascular Medicine*, Frontiers Media SA, v. 8, p. 768419, 2022.

MIIKKULAINEN, R. et al. Chapter 15 - evolving deep neural networks. In: KOZMA, R. et al. (Ed.). *Artificial Intelligence in the Age of Neural Networks and Brain Computing*. Academic Press, 2019. p. 293–312. ISBN 978-0-12-815480-9. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128154809000153>.

PASZKE, A. t. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2024. <https://pytorch.org/>. Acesso em: 10 jul. 2025.

RAISSI, M.; PERDIKARIS, P.; KARNIADAKIS, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, Elsevier, v. 378, p. 686–707, 2019.

REGAZZONI, F. et al. Learning the intrinsic dynamics of spatio-temporal processes through latent dynamics networks. *Nature Communications*, Nature Publishing Group UK London, v. 15, n. 1, p. 1834, 2024.

REIS, R. F. et al. A personalized computational model of edema formation in myocarditis based on long-axis biventricular mri images. *BMC bioinformatics*, Springer, v. 20, p. 1–11, 2019.

REIS, R. F. et al. An hydro-mechanical model of edema formation applied to bacterial myocarditis. In: IEEE. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. [S.l.], 2018. p. 1418–1424.

REIS, R. F. et al. On the mathematical modeling of inflammatory edema formation. *Computers & Mathematics with Applications*, Elsevier, v. 78, n. 9, p. 2994–3006, 2019.

SHAFI, O. et al. Demystifying tensorrt: Characterizing neural network inference engine on nvidia edge devices. In: IEEE. *2021 IEEE International Symposium on Workload Characterization (IISWC)*. [S.l.], 2021. p. 226–237.

SOMPAYRAC, L. M. *How the Immune System Works*. 4th. ed. [S.l.]: Blackwell Publising, 2010.

VICECONTI, M. et al. Credibility of in silico trial technologies—a theoretical framing. *IEEE journal of biomedical and health informatics*, IEEE, v. 24, n. 1, p. 4–13, 2019.

WANG, S.; YU, X.; PERDIKARIS, P. When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, v. 449, p. 110768, 2022. ISSN 0021-9991. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0021999121006630X>.

WANG, Y.-W.-Y. et al. Global, regional, and national burdens of myocarditis, 1990–2019: systematic analysis from gbd 2019: Gbd for myocarditis. *BMC Public Health*, Springer, v. 23, n. 1, p. 714, 2023.

WERNECK, Y. B. *Pinn-Torch: Repositório de códigos-fonte*. 2025. <https://github.com/ybwerneck/Pinn-Torch>. Acesso em: 10 jul. 2025.

WU, M. et al. Identification of necroptotic biomarkers associated with immune microenvironment in sepsis based on the protein–protein interaction network and machine learning. *Clinica Chimica Acta*, v. 577, p. 120489, 2025. ISSN 0009-8981. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0009898125003687>.

ZHANG, C. et al. Understanding deep learning requires rethinking generalization. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2017.

ZOHDI, T. Machine-learning and digital-twins for rapid evaluation and design of injected vaccine immune-system responses. *Computer Methods in Applied Mechanics and Engineering*, v. 401, p. 115315, 2022. ISSN 0045-7825. A Special Issue on Computational Modeling and Simulation of Infectious Diseases. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0045782522004169>.